ASURAT: functional annotation-driven unsupervised clustering of single-cell transcriptomes

Keita Iida[1,*], Jumpei Kondo[2,3], Johannes Nicolaus Wibisana[1], Masahiro Inoue[3], Mariko Okada[1,4]

[1] Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

[2] Division of Health Sciences, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan

[3] Department of Clinical Bio-resource Research and Development, Graduate School of Medicine Kyoto University, Kyoto 606-8501, Japan

[4] Center for Drug Design and Research, National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka, 567-0085, Japan

[*] Correspondence: kiida@protein.osaka-u.ac.jp

Abstract

Motivation: Single-cell RNA sequencing (scRNA-seq) analysis reveals heterogeneity and dynamic cell transitions. However, conventional gene-based analyses require intensive manual curation to interpret the biological implications of computational results. Hence, a theory for efficiently annotating individual cells is necessary.

Results: We present ASURAT, a computational pipeline for simultaneously performing unsupervised clustering and functional annotation of disease, cell type, biological process, and signaling pathway activity for single-cell transcriptomic data, using correlation graph-based decomposition of genes based on database-derived functional terms. We validated the usability and clustering performance of ASURAT using scRNA-seq datasets for human peripheral blood mononuclear cells, which required fewer manual curations than existing methods. Moreover, we applied ASURAT to scRNA-seq and spatial transcriptome datasets for small cell lung cancer and pancreatic ductal adenocarcinoma, identifying previously overlooked subpopulations and differentially expressed genes. ASURAT is a powerful tool for dissecting cell subpopulations and improving biological interpretability of complex and noisy transcriptomic data.

Availability: A GPLv3-licensed implementation of ASURAT is on GitHub (https://github.com/keita-iida/ASURAT).

Introduction

Single-cell RNA sequencing (scRNA-seq) has deepened our knowledge of biological complexity in terms of heterogeneity and dynamic transition of cell populations in a variety of phenomena, and this knowledge has immense potential for elucidating the regulatory principles underlying our body plans (La Manno, et al., 2018). scRNA-seq has been widely used to improve the molecular understanding of malignant cells in lymphoma (Zhang, et al., 2019), intra- and intertumoral heterogeneity in drug-treated cancer populations (Stewart, et al., 2020), ligand-receptor interaction in tumor immune microenvironments (Chen, et al., 2020), and the effects of viral infection on immune cell populations (Devitt, et al., 2019). Various clustering methods based on gene expression similarity have been proposed (Pasquini, et al., 2021) and applied to annotate cell types (Kim, et al., 2020). However, conventional gene-based analyses require intensive manual curation to annotate clustering results; hence, efficient and unbiased interpretation of single-cell data remains challenging (Andrews, et al., 2021; Aran, et al., 2019; Gao, et al., 2019; Kiselev, et al., 2019; Lahnemann, et al., 2020).

Conventionally, single-cell transcriptomes are analyzed and interpreted by means of unsupervised clustering followed by manual curation of marker genes chosen from a large number of differentially expressed genes (DEGs) (Andrews, et al., 2021; Lahnemann, et al., 2020). Here, manual curations are based on literature searches of biological functions of DEGs. Today, several computational tools for semi-automated cell type and marker gene inference based on clustering results are available to assist manual annotation, as detailed in the review by Pasquini *et al.* (2021). However, this is often difficult because a single gene is generally multifunctional and therefore associated with multiple biological function terms (Cancer Genome Atlas Research, et al., 2017). In cancer transcriptomics, this difficulty is exacerbated by the complex interdependence between disease-related biomarker genes and their heterogeneous expressions, which are associated with numerous biological function terms.

A possible solution is to realize cell clustering and biological interpretation at the same time. Recently, reference component analysis (RCA), which is used for accurate clustering of single-cell transcriptomes along with unbiased cell-type annotation based on similarity to reference transcriptome panels (Li, et al., 2017). Yet, these methods require the transcriptomic data of well-characterized reference cells as learning datasets, which might not always be available. Another approach is using supervised classification (Gao, et al., 2019) combined with gene set enrichment analysis, incorporating biological knowledge and functions; hence, it may improve the interpretability over signature gene-based approaches, which place sole emphasis on individual roles of genes (Fan, et al., 2016). Despite these advances, we still lack a prevailing theory leveraging this information at the single-cell level.

To overcome the aforementioned limitations, a method providing simultaneous interpretation of biological function and classification of the cells is needed for single-cell analysis. Thus, we propose an original computational pipeline named ASURAT (functional annotation-driven unsupervised clustering of single-cell transcriptomes), which simultaneously performs unsupervised cell clustering and biological interpretation in terms of cell type, disease, biological process, and signaling pathway activity. In this study, we demonstrate the clustering performance of ASURAT using standard scRNA-seq and spatial transcriptome (ST) datasets for human peripheral blood mononuclear cells (PBMCs), small cell lung cancer (SCLC), and pancreatic ductal adenocarcinoma (PDAC), respectively. We show that ASURAT can greatly improve functional understanding of single-cell transcriptomes, adding a new layer of biological interpretability to conventional gene-based analyses.

Methods

Overview of ASURAT workflow

ASURAT was developed to simultaneously cluster and interpret single-cell transcriptomes using functional gene sets (FGSs) (**Figure 1**), and it was implemented in the R programming language. FGSs are collected from knowledge-based databases (DBs) for disease, cell type, biological process, and signaling pathway activity, such as Disease Ontology (DO) (Yu, et al., 2015), Cell Ontology (CO) (Diehl, et al., 2016), Gene Ontology (GO) (Yu, et al., 2012), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) by implementing R packages such as DOSE (version 3.16.0), ontoProc (version 1.12.0), clusterProfiler (version 3.18.0), and

KEGGREST (version 1.30.0), respectively (**Figure 1b**). Then, ASURAT created multiple biological terms using single-cell transcriptome data and the FGSs (**Figure 1c**, Supplementary Note 1). We called such new biological terms signs. Finally, ASURAT created a sign-by-sample matrix (SSM), in which rows and columns stand for signs and samples (cells), respectively (**Figure 1c**). SSM is analogous to a read count table, where the rows represent signs with biological meaning instead of individual genes and that the values contained are sign scores instead of read counts. By analyzing SSMs, individual cells can be characterized by various biological terms (**Figure 1d**).

Sign

Let A be a read count table of size $p \times n$ from single-cell transcriptomic data, whose rows and columns mean p genes, represented by $\Omega = \{1, 2, \cdots, p\}$, and n cells, respectively, and R a "relation" (e.g., correlation matrix) among Ω. Let $\mathcal{F} = \{(T_k, \Omega_k) | k = 1, 2, \cdots, q\}$ be a set of ordered pairs, where T_k and $\Omega_k \in 2^\Omega$ (2^Ω is a power set of Ω) are biological description and the FGS, respectively. Consider an R-dependent representation $\Omega_k = \bigcup_{j=1}^{m_k} \Omega_k^{(j)}$, where m_k is an integer, for $k = 1, 2, \cdots, q$. Then, the triplet $(T_k, \Omega_k^{(j)}, R)$ is termed a sign, in particular (T_k, Ω_k, R) a parent sign. Our definition is inspired by Saussure's semiology as described in the early 20th century. According to Maruyama (2008), the original notion of a *signe* is a segment of a thing of interest, which is created by an arbitrary decomposition based on its relationships. For example, a rainbow is a continuum of varying light input, from which we can see distinct colors of red, yellow, green, and blue by our subjective decomposition based on their spectral relationships (Couper, 2015).

Correlated gene set

Let $R = (r_{i,j})$ be a correlation matrix of size $p \times p$ defined by A and a certain measure (e.g., Pearson's measure), whose diagonal elements are 1. Let α and β be positive and negative constants satisfying $0 < \alpha \leq 1$ and $-1 \leq \beta < 0$, respectively. Let us arbitrarily fix $(T_k, \Omega_k) \in \mathcal{F}$ and consider the following subsets of Ω_k:

$$U_k(\alpha) = \{i \in \Omega_k | \exists j \in \Omega_k \text{ such that } r_{i,j} \geq \alpha, i \neq j\},$$
$$V_k(\beta) = \{i \in \Omega_k | \exists j \in \Omega_k \text{ such that } r_{i,j} \leq \beta, i \neq j\},$$

$$W_k(\alpha, \beta) = U_k(\alpha) \cup V_k(\beta).$$

Hereinafter we omit the arguments α and β for simplicity. Let us denote $\Omega_k^{(\mathrm{w})} = \Omega_k - W_k$. If V_k is not empty, represent each element of W_k as a point in the Euclidean space spanned by the row vectors of R and decompose W_k into two disjoint subsets by Partitioning Around Medoids (PAM) clustering (Schubert and Rousseeuw, 2019), that is $W_k = \Omega_k^{(\mathrm{s})} \cup \Omega_k^{(\mathrm{v})}$. Otherwise, if V_k is empty, let $\Omega_k^{(\mathrm{s})} = U_k$ and $\Omega_k^{(\mathrm{v})} = \phi$ (empty). Thus Ω_k is decomposed into three parts as follows:

$$\Omega_k = \Omega_k^{(\mathrm{s})} \cup \Omega_k^{(\mathrm{v})} \cup \Omega_k^{(\mathrm{w})}. \tag{1}$$

Let $\mu_k^{(\mathrm{s})}$ (resp. $\mu_k^{(\mathrm{v})}$) be the mean of off-diagonal elements of R for $\Omega_k^{(\mathrm{s})}$ ($\Omega_k^{(\mathrm{v})}$), and assume $\mu_k^{(\mathrm{s})} \geq \mu_k^{(\mathrm{v})}$ without loss of generality. If $\mu_k^{(\mathrm{s})} \geq \alpha$, then $\Omega_k^{(\mathrm{s})}$, $\Omega_k^{(\mathrm{v})}$, and $\Omega_k^{(\mathrm{w})}$ are termed strongly, variably, and weakly correlated gene sets, respectively, which are hereafter abbreviated as SCG, VCG, and WCG. Otherwise, correlated gene sets cannot be defined for T_k.

For any given (T_k, Ω_k, R) the genes should strongly and positively correlate within each of the $\Omega_k^{(\mathrm{s})}$ and $\Omega_k^{(\mathrm{v})}$, while they should negatively correlate between $\Omega_k^{(\mathrm{s})}$ and $\Omega_k^{(\mathrm{v})}$. Thus, one can hypothesize that SCG and VCG are predominantly associated with T_k, which may aid interpretation of biological meanings of corresponding signs. **Figure 2** shows that the SCG and VCG include *KRT18* and *ASCL1*, which respectively have negative and positive contributions for lung small cell carcinoma. Thus, we interpret that $(T_k, \Omega_k^{(\mathrm{s})}, R)$ and $(T_k, \Omega_k^{(\mathrm{v})}, R)$ for DOID 5409 relate positively and negatively with this cell type, respectively.

Though simpler methods decomposing correlation graphs exist, such as one-shot PAM clustering (Schubert and Rousseeuw, 2019), hierarchical clustering and tree cutting (Murtagh and Legendre, 2014), principal component analysis (PCA)-based methods (Hyvarinen, 1999), and several graph statistical approaches (Blondel, et al., 2008; Bodenhofer, et al., 2011), we found that our VCG definition is critical for clustering cells. In fact, we tried replacing our decomposition method (1) with one-shot PAM clustering, but the results frequently exhibited deteriorated performance because both VCG and WCG (obtained from the one-shot clustering) included many weakly correlated genes.

Sign-by-sample matrix

Let $A = (a_{i,j})$ be a gene-by-cell matrix of size $p \times n$ from a single-cell transcriptomic data, whose entries stand for normalized-and-centered gene expression levels. For simplicity, let us assume that functional gene sets Ω_k can be decomposed into non-empty $\Omega_k^{(\mathrm{s})}$, $\Omega_k^{(\mathrm{v})}$, and $\Omega_k^{(\mathrm{w})}$, for $k = 1, 2, \cdots, q$. Let $B^{(x)}$, $x \in \{\mathrm{s}, \mathrm{v}, \mathrm{w}\}$, be matrices of size $q \times n$, whose entries $b_{k,j}^{(x)}$ are defined as follows:

$$b_{k,j}^{(x)} = \frac{1}{|\Omega_k^{(x)}|} \sum_{i \in \Omega_k^{(x)}} a_{i,j},$$

where $|\Omega_k^{(x)}|$ stands for the number of elements in $\Omega_k^{(x)}$. Additionally, let $C^{(x)}$, $x \in \{\mathrm{s}, \mathrm{v}\}$, be $q \times n$ matrices as follows:

$$C^{(x)} = \omega^{(x)} B^{(x)} + \left(1 - \omega^{(x)}\right) B^{(\mathrm{w})}, \qquad (2)$$

where $\omega^{(x)}$, $0 \leq \omega^{(x)} \leq 1$, are weight constants. Here $C^{(\mathrm{s})}$ and $C^{(\mathrm{v})}$ are termed sign-by-sample matrices (SSMs) for SCG and VCG, respectively, and the entry $c_{k,j}^{(x)}$ a sign score of the kth sign and jth sample (cell). By vertically concatenating SSMs for SCGs and VCGs, we created a single SSM. Note that ensemble means of sign scores across cells are zeros because SSMs are derived from the centered gene expression matrix A.

Unsupervised clustering of sign-by-sample matrices

One focus of analyzing SSMs is to cluster cells and find significant signs (**Figure 1d**), where "significant" means that the sign scores, i.e., the entries of (2), are specifically upregulated or downregulated at the cluster level. It should be noted that significant signs are analogous to DEGs but bear biological meanings. Here, naïve usages of statistical tests and fold change analyses should be avoided because the row vectors of

SSMs are centered. Hence, we propose a nonparametric separation index, which quantifies the extent of separation between two sets of random variables (Supplementary Note 2). To cluster cells, we used two strategies. The first is unsupervised clustering, such as PAM, hierarchical, and graph-based clustering. The second is a method of extracting a continuous tree-like topology using diffusion map (Coifman and Lafon, 2006), followed by allocating cells to different branches of the data manifolds (Parra, et al., 2019). Choosing an appropriate strategy depends on the biological context, but the latter is usually applied to developmental processes or time-course experimental data, which are often followed by pseudotime analyses.

Results

Clustering single-cell transcriptomes of peripheral blood mononuclear cells

To validate the usability and clustering performance of ASURAT in comparison with the existing methods, we analyzed two public scRNA-seq datasets, namely the PBMC 4k and 6k datasets (Supplementary Note 3), in which the cell types were inferred using computational tools based on prior assumptions (Cao, et al., 2020). We first excluded low-quality genes and cells and attenuated technical biases with respect to zero inflation and variation of capture efficiencies between cells using bayNorm (Tang, et al., 2020) (Supplementary Note 4). The resulting read count tables were supplied to ASURAT and four other methods: scran (version 1.18.7) (Lun, et al., 2016), Seurat (version 4.0.2) (Hao, et al., 2021), Monocle 3 (version 1.0.0) (Trapnell, et al., 2014), and SC3 (version 1.18.0) (Kiselev, et al., 2017). To infer existing cell types and the population ratios in the PBMC 4k and 6k datasets, we implemented the existing methods using settings close to the default ones, performed cell clustering, and annotated each cluster by manually investigating DEGs based on the false discovery rates (FDRs)$< 10^{-99}$ (**Figure 3a**, Supplementary Note 5). When using ASURAT, we performed unsupervised cell clustering and semi-automatic annotation based on SSMs for CO, GO, and KEGG.

Among all the existing methods, Seurat and Monocle 3 could robustly reproduce most blood cell type labels, as inferred by Cao *et al.* (2020), while scran and SC3 output many unspecified cells (**Figure 3c**). We found that the Seurat pipeline, followed by manual annotations based only on a couple of DEGs, provided comparable population ratios with previous results (Cao, et al., 2020). However, it was quite laborious to manually select marker genes from numerous DEGs (**Figure 3a**), which tend to increase in terms of the number of cells as well as significance levels. Based on the clustering results of Seurat, we assigned the labels (i) T cell, (ii) monocyte, (iii) B cell, and (iv) NK or NKT cell to the cells in PBMC 4k (resp. PBMC 6k) by finding marker genes from (i) 57 (114), (ii) 102 (148), (iii) 49 (33), and (iv) 32 (35) DEGs, respectively. To avoid such a laborious process, it is possible to implement automatic annotation tools based on the calculated DEGs, such as by using scCATCH (version 2.1) (Shao, et al., 2020). Nevertheless, population ratios inferred by Seurat with scCATCH were less consistent than those by Seurat with manual annotations (**Figure 3c**).

ASURAT simultaneously performed unsupervised cell clustering and biological interpretation leveraging all defined FGSs, without relying on DEGs (**Figure 3a**). We identified five cell type labels, with none remaining unspecified (**Figure 3b, c**, Figure S2). The population ratios were approximately consistent with the reported values (Cao, et al., 2020), except for the small dendritic cell population possibly included in PBMCs (Villani, et al., 2017; Wagner, 2018). Such a small discrepancy was unavoidable, because Cao *et al*. (2020) used author-defined DEGs and preselected cell types to identify the most preferable ones. Unexpectedly, the clustering results using SSMs for GO and KEGG also showed well-separated clusters in two-dimensional Uniform Manifold Approximation and Projection (UMAP) (McInnes and Healy, 2018) spaces (Figure S3), indicating that the functional states of cells are also heterogeneous with respect to biological process and signaling pathway activity. These results demonstrate that ASURAT can perform robust clustering for single-cell transcriptomes.

Clustering a small cell lung cancer single-cell transcriptome

SCLC tumors undergo a transition from chemosensitivity to chemoresistance states against platinum-based therapy (Stewart, et al., 2020). Stewart *et al*. (2020) analyzed scRNA-seq data obtained from circulating tumor cell-derived xenografts generated from treatment-naïve lung cancer patients, cultured them with vehicle or cisplatin treatments, and reported that the gene expression profiles of the platinum-resistant tumors were more heterogeneous than those of platinum-sensitive tumors. However, the mechanism behind chemoresistance remains unclear, partly because transcriptional heterogeneity is affected by physiological states of cells such as pathological states (Stewart, et al., 2020), cell cycle (Dominguez, et al., 2016), and metabolic processes (Jalili, et al., 2021), which cannot be readily identified by conventional marker gene-based analyses alone. To better understand SCLC subtypes in chemoresistant tumors, we applied Seurat and ASURAT to the published SCLC scRNA-seq data (Supplementary Note 3) (Stewart, et al., 2020).

First, we investigated the expression levels of known SCLC marker genes (Ireland, et al., 2020), namely *ASCL1*, *NEUROD1*, *YAP1*, and *POU2F3* and confirmed that almost all of the cells are of the *ASCL1* single-positive subtype (Figure S4), which is consistent with the previous report (Stewart, et al., 2020). After quality controls, the data were normalized by bayNorm (Tang, et al., 2020) and the resulting read count table was supplied to the workflows of Seurat and ASURAT (Supplementary Note 4, 6). To investigate molecular subtypes and potential resistance pathways, we clustered the single-cell transcriptome and inferred a cell cycle phase for each cell using Seurat (Hao, et al., 2021), as shown in the UMAP spaces (**Figure 4e**). We found that the cell populations assigned to G1, S, and G2M phases are sequentially distributed in the UMAP space, indicating that the clustering results are considerably affected by the cell cycle. Then, we identified DEGs for each cluster (Group 1, 2, and 3) and performed KEGG enrichment analysis using clusterProfiler (Yu, et al., 2012), but the chemoresistance terms were not primarily enriched (**Figure 4f**).

Subsequently, to investigate functional heterogeneities in SCLCs, we used ASURAT to create SSMs using DO, GO, and KEGG. Based on the SSM for DO, we performed a dimensionality reduction using diffusion map (Coifman and Lafon, 2006), which showed a tree-like topology. Then, we defined a pseudotime along the branches and clustered the single-cell transcriptome using MERLoT (Parra, et al., 2019) (**Figure 4c**). Based on pseudotime analysis, we revealed that sign scores for platinum drug resistance (path:hsa01524_S) and PD-L1 expression-mediated immunosuppression (path:hsa05235_S) were upregulated in clusters 2 and 3, respectively. In addition, sign scores for intracellular protein transport (GO:0006886_S), with an FGS including the

SCLC malignancy marker *CD24* (Kristiansen, et al., 2003), was upregulated in cluster 1 (**Figure 4d**). We noticed that sign scores for hematopoietic system disease (DOID:74_S) were moderately increased in cluster 1 (separation index~0.38), which was supported by a previous work reporting that hematopoietic cancers are similar to SCLCs in terms of gene expression profiles and drug sensitivities (Balanis, et al., 2019). Although the SCLC molecular subtypes have been extensively studied (Chen, et al., 2019; Ireland, et al., 2020; Schwendenwein, et al., 2021; Wooten, et al., 2019; Yatabe, 2020), data regarding the functional subtypes of *ASCL1*-positive SCLC remain limited. To identify *de novo* SCLC subtypes, future work will validate our clustering results.

Finally, we vertically concatenated all the SSMs, cell cycle phases, and expression matrices to characterize individual cells from multiple biological aspects, as shown by the heatmaps along with the clustering result of ASURAT (**Figure 4a, b**). As shown, we were able to simultaneously perform unsupervised clustering and biological interpretation of single-cell transcriptomes. Moreover, we added a layer of DEGs using multiple Mann-Whitney *U* tests (**Figure 4a**), showing that most DEGs had been previously overlooked (Chen, et al., 2019; Ireland, et al., 2020; Schwendenwein, et al., 2021; Wooten, et al., 2019; Yatabe, 2020). Taken together, we provide a novel clue for the clinical improvements for relapsed SCLC tumors.

Clustering a pancreatic ductal adenocarcinoma spatial transcriptome

Moncada *et al*. (2020) analyzed scRNA-seq and ST data obtained from PDAC patients (Moncada, et al., 2020) and reported that cancer and non-cancer cells are spatially distributed in the distinct tissue regions of the primary PDAC tumors, and that PDAC cells are accompanied by inflammatory fibroblasts. Since the cellular resolutions of the STs were estimated to be 20–70 cells per ST spot, which is far lower than that of scRNA-seq, computational methods have been proposed to predict existing cell types by integrating ST and scRNA-seq datasets (Elosua-Bayes, et al., 2021; Moncada, et al., 2020). Here, we aimed to dissect ST data and compare the annotation results of ASURAT with those of Seurat by using ST (PDAC-A ST1) and scRNA-seq (PDAC-A inDrop from 1 to 6) datasets (Moncada, et al., 2020) (Supplementary Note 3).

First, we combined all the scRNA-seq datasets after confirming that there were minimal batch effects (Figure S5). Then, the ST and scRNA-seq data were normalized by bayNorm (Tang, et al., 2020) (Supplementary Note 4) and the resulting read count tables were supplied to Seurat. To cluster the ST with reference to the scRNA-seq data, we performed canonical correlation analysis (CCA)-based data integration of Seurat (**Figure 5a**), followed by an unsupervised clustering of the integrated transcriptome using Seurat functions, which is shown in UMAP spaces (**Figure 5b**) and the tissue image (**Figure 5c**). Unexpectedly, batch effects were not corrected between ST and scRNA-seq datasets after data integration; nevertheless, the inferred cancer and non-cancer regions were approximately consistent with previously annotated histological regions (Elosua-Bayes, et al., 2021; Moncada, et al., 2020), wherein several marker genes such as *REG1A*, *S100A4* and *TM4SF1*, and *CELA2A* were identified as DEGs for clusters 2, 3, and 5, respectively (FDRs$< 10^{-80}$, Mann-Whitney U tests).

Next, we input the ST and scRNA-seq integrated transcriptome into ASURAT workflow. To investigate complex PDAC tissues, we created SSMs using DO, CO, GO, and KEGG, as well as CellMarker (Zhang, et al., 2019) and MSigDB (Subramanian, et al., 2005). Based on the SSM for GO, which was computed from the integrated transcriptome, we performed a dimensionality reduction using PCA and clustered the SSM by *k*-nearest neighbor (KNN) graph generation and the Louvain algorithm, which is shown in UMAP spaces (**Figure 5d**) and the tissue image (**Figure 5e**). Remarkably, ASURAT was able to remove the aforementioned batch effects and infer the spots we suspect as atypical region which might be a normal pancreas involved in cancer (Figure 5e left bottom).

To further investigate cell states in these spots, we computed all the sign scores across the tissue (Figure S6). We found that the sign scores for PDAC (DOID:3498_S), which has an FGS including PDAC markers such as *S100P* and *MMP1*, were increased in the ST spots approximately matching the reported PDAC region (Moncada, et al., 2020), while those for transcriptional misregulation in cancer (path:hsa05202_S) and microRNAs in cancer (path:hsa05206_S) were increased both in the previously annotated PDAC spots and the newly predicted atypical spots (**Figure 5f**). These newly predicted spots were also annotated by a sign for Th17 cell differentiation (path:hsa04659_S), suggesting tumor-associated inflammation or antitumor immunity through intercellular communications between Th17 and cancer cells (Muller-Hubenthal, et al., 2009), which remains to be elucidated in PDAC (Liu, et al., 2019).

It is reported that in more than 90% of PDAC cases, *KRAS* is mutated at the G domain of the 12th residue (Ischenko, et al., 2021; Luchini, et al., 2020). Hence, we speculated that it might be possible to validate our clustering results of cancer and non-cancer spots by comparing the frequencies of KRAS mutations using ST data. Unfortunately, we were unable to detect any read mapped to the specific reported region, possibly owing to the shallow read depth and inherent 3′ bias present in the data. We hope that simultaneous genetic and transcriptional profiling can address this problem in the future (Lee, et al., 2020).

Discussion

We have developed ASURAT, a novel computational pipeline for simultaneous cell clustering and biological interpretation using FGSs. ASURAT begins by performing a correlation graph-based decomposition of FGS to define multiple biological terms, termed signs. ASURAT then transforms scRNA-seq data into an SSM, whose rows and columns stand for signs and samples, respectively. This SSM plays a key role in characterizing individual cells by various biological terms. Applying ASURAT to several scRNA-seq and spatial transcriptome datasets for PBMCs, SCLC, and PDAC, we robustly reproduced the previously reported blood cell types, identified putative subtypes of chemoresistant SCLC, and identified distinct regions within the PDAC tissue.

Conventionally, single-cell transcriptomes are analyzed and interpreted by means of unsupervised clustering followed by manual curation of marker genes chosen from a large number of DEGs, which has been a common bottleneck of gene-based analyses (Andrews, et al., 2021; Aran, et al., 2019; Gao, et al., 2019). The statistical significance of individual genes, typically defined by *p*-value or fold change, is dependent on clustering results, which are also affected by various physiological states of cells

(Dominguez, et al., 2016; Jalili, et al., 2021). Here, we expect that ASURAT provides an alternative approach using FGSs and demonstrates superior performance for identifying functional subtypes even within a fairly homogeneous population such as isolated cancer cells. In practice, complemental usages of ASURAT and existing methods (Butler, et al., 2018; La Manno, et al., 2018) will provide more comprehensive understanding of single-cell and spatial transcriptomes, helping us shed light on putative transdifferentiation of neuroendocrine cancers (Balanis, et al., 2019; Kubota, et al., 2020), intercellular communication in tumor immune microenvironments (Maynard, et al., 2020), and virus infection on immune cell populations (Devitt, et al., 2019).

In omics data analyses, knowledge-based DBs are used to interpret computational results: GO, KEGG pathway, and motif enrichment analyses are often used for transcriptomic and epigenomic analyses (McLeay and Bailey, 2010; Mootha, et al., 2003; Reimand, et al., 2019). In contrast, we propose a unique analytical workflow, in which such DBs are used for simultaneous clustering and biological interpretation by defining signs from single-cell transcriptome data and FGSs. This framework is potentially applicable to any multivariate data with variables linked with annotation information. We can also find such datasets in studies of T cell receptor sequencing (De Simone, et al., 2018; Rempala, et al., 2011) along with a pan-immune repertoire (Zhang, et al., 2020). We anticipate that ASURAT will make it possible to identify various inter-sample differences among T cell receptor repertoires in terms of cellular subtype, antigen-antibody interaction, genetic and pathological backgrounds.

Finally, future challenges in data-driven mathematical analysis are worth noting. Since ASURAT can create multivariate data (i.e., SSMs) from multiple signs, ranging from cell types to biological functions, it will be valuable to consider graphical models of signs, from which we may infer conditional independence structures. A non-Gaussian Markov random field theory (Morrison, et al., 2017) is one of the most promising approaches to this problem, but it requires quite a large number of samples for achieving true graph edges (Morrison, et al., 2017). As available data expand in size and diversity, biological interpretation will become increasingly important. Hence, future work should improve methods for prioritizing biological terms more efficiently than manual screening. We hope ASURAT will greatly facilitate our intuitive understanding of various biological data and open new means of general functional annotation-driven data analysis.

Acknowledgements

We thank Dr. Takeya Kasukawa for comments that improved the analysis pipeline.

Funding

K.I. was supported by JSPS KAKENHI Grant No. 20K14361. K.I., J.K., and M.I. were supported by Shin Bunya Kaitaku Shien Program of Institute for Protein Research, Osaka University. M.O. was supported by JSPS KAKENHI Grant No. 17H06299, 17H06302, and 18H04031, JST-Mirai program No. JPMJMI19G7, and JST CREST Program No. JPMJCR21N3. M.O. and M.I. were supported by P-CREATE, Japan Agency for Medical Research and Development. M.O. and K.I. were supported by JST Moonshot R&D Grant Number JPMJMS2021.

Data Availability statement

The PBMCs datasets are available in the 10x Genomics repository at https://support.10xgenomics.com/single-cell-gene-expression/datasets. The SCLC and PDAC datasets are available in Gene Expression Omnibus with accession codes GSE138474 (GSM4104164) and GSE111672 (GSM3036909, GSM3036910, GSM3036911, GSM3405527, GSM3405528, GSM3405529, and GSM3405530), which are referenced in (Stewart, et al., 2020) and (Moncada, et al., 2020), respectively.

Conflicts of Interest: none declared.

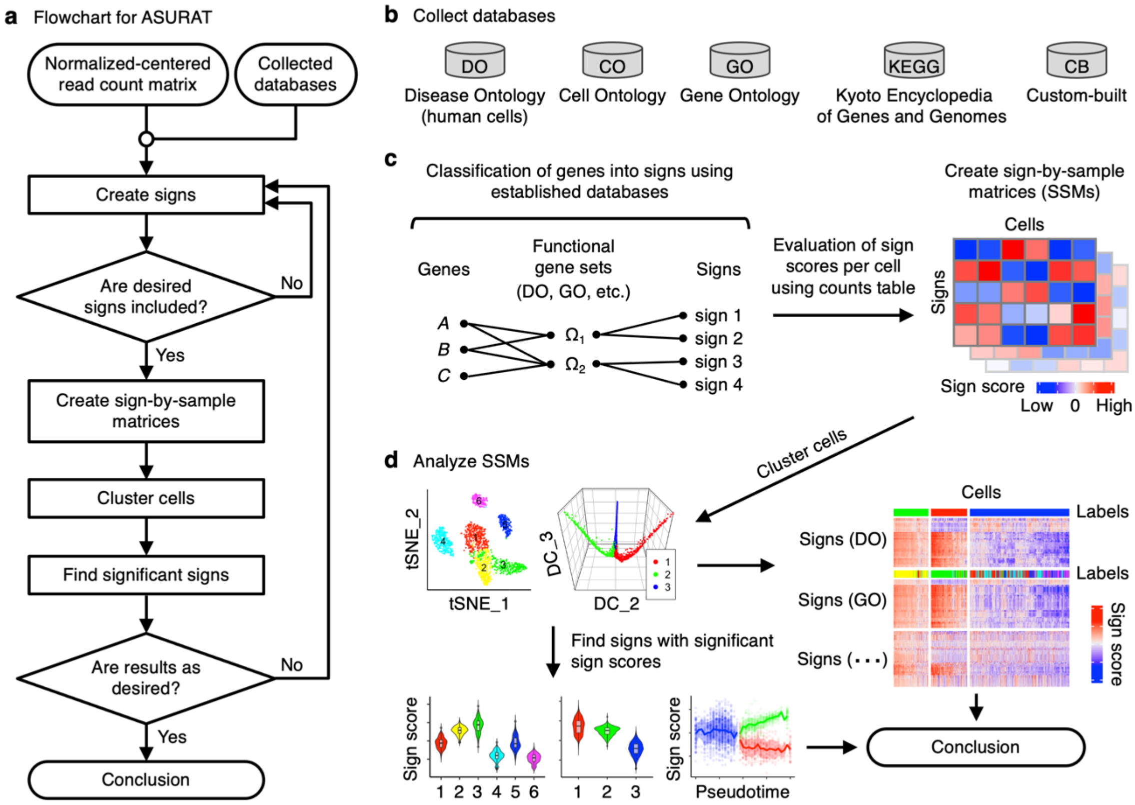


Figure 1. Workflow of ASURAT. (a) Flowchart of the procedures. (b) Collection of knowledge-based data-bases (DBs). (c) Creation of sign-by-sample matrices (SSMs) from normalized-and-centered read count table and the collected DBs. (d) Analysis of SSMs to infer diseases, cell types, biological processes, and signaling pathway activities.

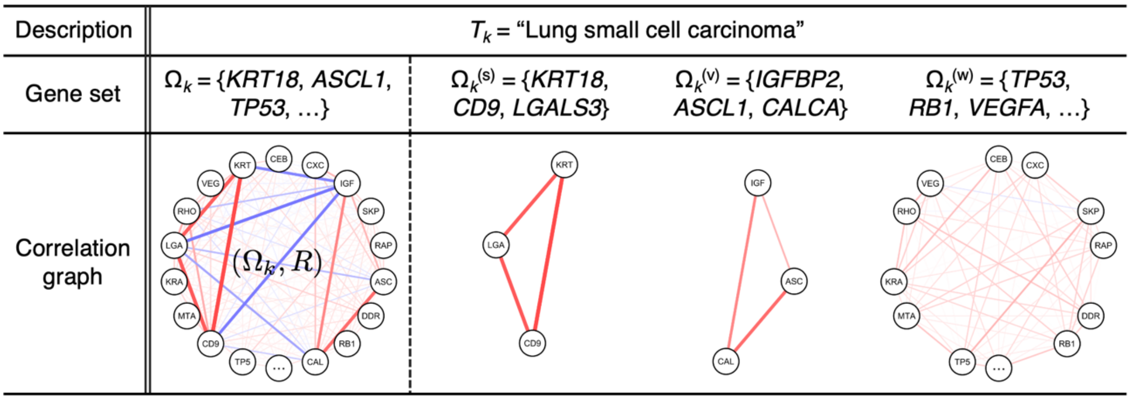

Description
T_k = "Lung small cell carcinoma"
Gene set
Ω_k = {KRT18, ASCL1, TP53, …}
$\Omega_k^{(s)}$ = {KRT18, CD9, LGALS3}
$\Omega_k^{(v)}$ = {IGFBP2, ASCL1, CALCA}
$\Omega_k^{(w)}$ = {TP53, RB1, VEGFA, …}
Correlation graph
(Ω_k, R)
KRT
CEB
CXC
VEG
IGF
RHO
SKP
LGA
RAP
KRA
ASC
MTA
DDR
CD9
RB1
TP5
…
CAL

Figure 2. Representation of correlation graph-based decomposition. From single-cell RNA sequencing data and a Disease Ontology (DO) term with DOID 5409, which concerns small cell lung cancer, three signs $(T_k, \Omega_k^{(j)}, R)$, $j \in \{\mathrm{s}, \mathrm{v}, \mathrm{w}\}$, were produced from their parent sign (T_k, Ω_k, R) by decomposing the correlation graph (Ω_k, R) into strongly, variably, and weakly correlated gene sets: $\Omega_k^{(\mathrm{s})}$, $\Omega_k^{(\mathrm{v})}$, and $\Omega_k^{(\mathrm{w})}$, respectively. Red and blue edges in correlation graphs indicate positive and negative correlations, respectively; color density indicates the strength of the correlation.

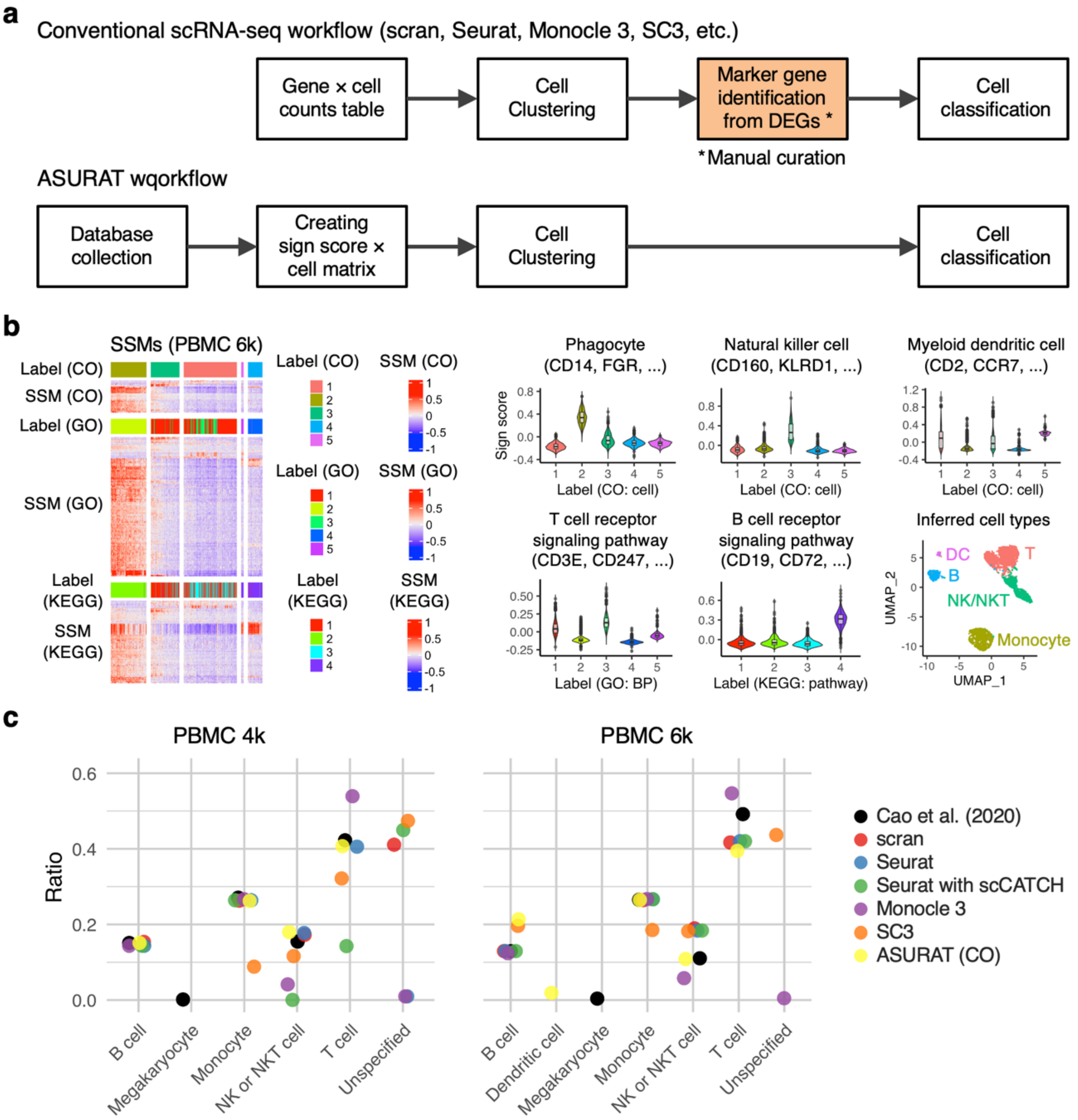
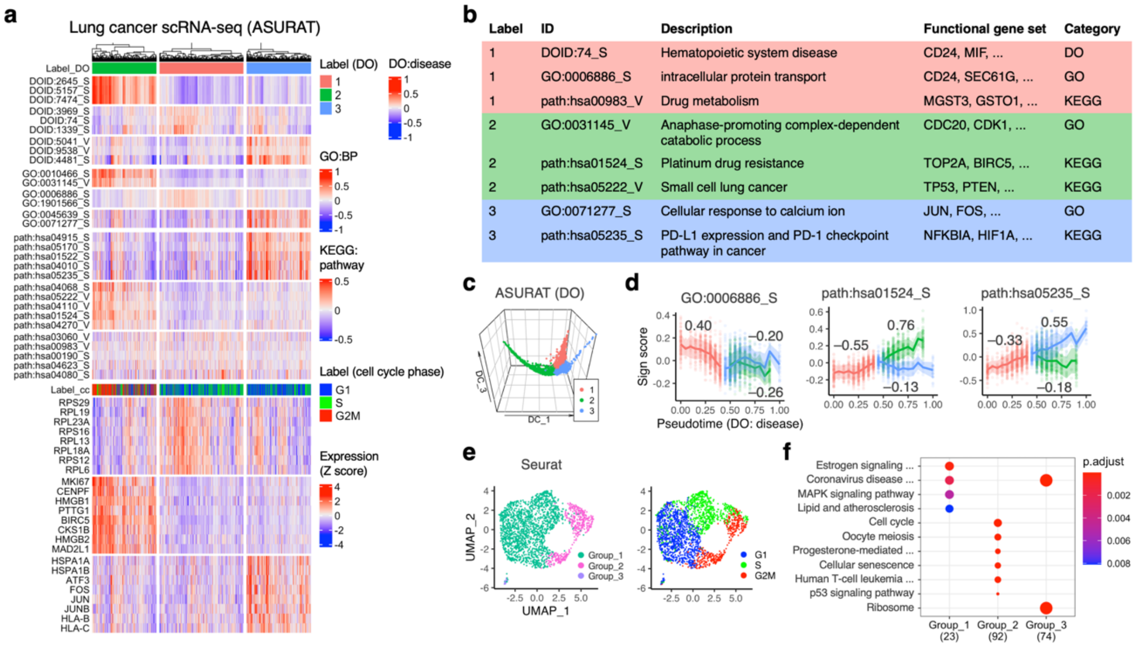


Figure 3. Clustering peripheral blood mononuclear cell (PBMC) single-cell transcriptomes. (a) Schematic illustration of conventional single-cell RNA sequencing and ASURAT workflows. (b) Identification of cell types in the PBMC 6k dataset from analyses of sign-by-sample matrices (SSMs) for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG). According to heatmaps and violin plots of representative signs and functional gene sets, T cell ("T"), B cell ("B"), NK or NKT cell ("NK/NKT"), monocyte, and dendritic cell ("DC") were identified as shown in Uniform Manifold Approximation and Projection (UMAP) plots.

(c) Population ratios in the PBMC 4k and 6k datasets predicted by seven different methods. DEG, differentially expressed gene.



Label	ID	Description	Functional gene set	Category
1	DOID:74_S	Hematopoietic system disease	CD24, MIF, ...	DO
1	GO:0006886_S	intracellular protein transport	CD24, SEC61G, ...	GO
1	path:hsa00983_V	Drug metabolism	MGST3, GSTO1, ...	KEGG
2	GO:0031145_V	Anaphase-promoting complex-dependent catabolic process	CDC20, CDK1, ...	GO
2	path:hsa01524_S	Platinum drug resistance	TOP2A, BIRC5, ...	KEGG
2	path:hsa05222_V	Small cell lung cancer	TP53, PTEN, ...	KEGG
3	GO:0071277_S	Cellular response to calcium ion	JUN, FOS, ...	GO
3	path:hsa05235_S	PD-L1 expression and PD-1 checkpoint pathway in cancer	NFKBIA, HIF1A, ...	KEGG

Figure 4. Clustering a single-cell transcriptome of small cell lung cancers. (a) Heatmaps showing (i) clustering results of ASURAT, (ii) sign scores of sign-by-sample matrices (SSMs) for Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), and (iii) scaled gene expression levels, which are concatenated vertically. Here, only the most significant signs and differentially expressed genes (DEGs) for ASURAT clusters are shown. (b) Representative signs from (a). (c) Diffusion map of the SSM for DO, projected onto the first three coordinates. (d) Sign scores for the indicated IDs along the pseudotime, in which the standard deviations are shown by the shaded area. The value on each plot stands for the separation index for a given group versus all the others. The clustering labels are consistent with those in (a) and (b). (e) Clustering results and cell cycle phases computed by Seurat. (f) KEGG pathway enrichment analysis based on DEGs for Seurat clusters in (e).

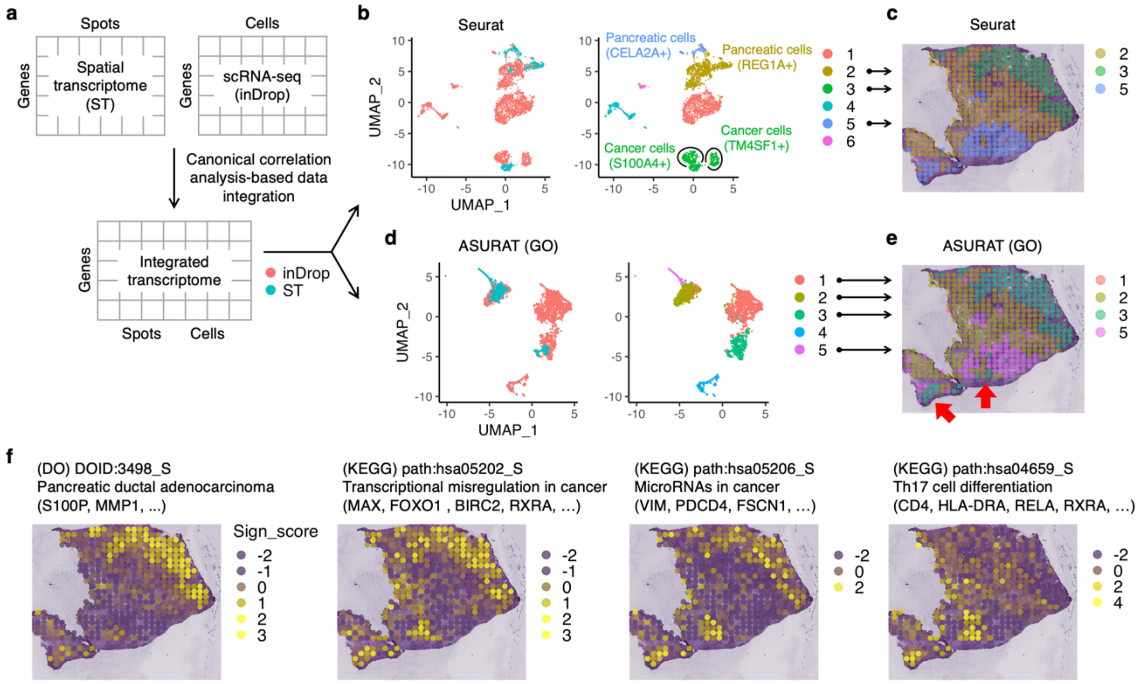


Figure 5. Clustering of spatial transcriptome (ST) data of pancreatic ductal adenocarcinoma (PDAC). (a) Canonical correlation analysis-based data integration of single-cell RNA sequencing (scRNA-seq) and ST datasets using Seurat. (b) Seurat unsupervised clustering based on the integrated data. Cells were manually labeled according to the indicated differentially expressed genes (DEGs) in Uniform Manifold Approximation and Projection (UMAP) plots. (e) ASURAT clustering result shown in the PDAC tissue, in which red arrows indicate the spots newly predicted as atypical region which might be a normal pancreas involved in cancer. (f) Profiles of sign scores in the PDAC tissue, predicting cancer and inflammation spots. DO, Disease Ontology. GO, Gene Ontology. KEGG, Kyoto Encyclopedia of Genes and Genomes.

Supplementary Materials

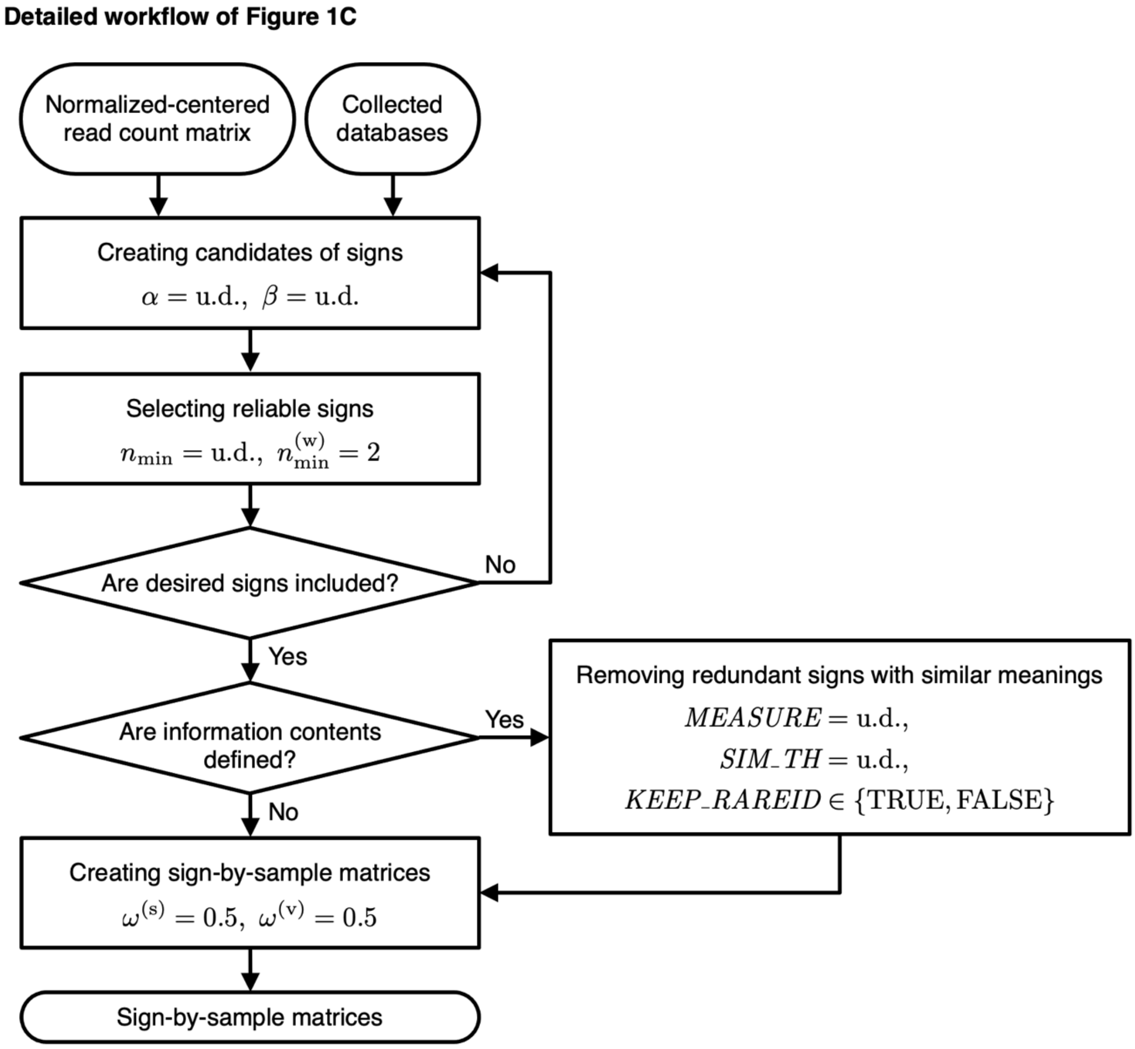

Detailed workflow of Figure 1C
Normalized-centered read count matrix
Collected databases
Creating candidates of signs
$\alpha = \text{u.d.}, \ \beta = \text{u.d.}$
Selecting reliable signs
$n_{\text{min}} = \text{u.d.}, \ n^{(\text{w})}_{\text{min}} = 2$
Are desired signs included?
No
Yes
Are information contents defined?
Yes
No
Removing redundant signs with similar meanings
$MEASURE = \text{u.d.},$
$SIM\_TH = \text{u.d.},$
$KEEP\_RAREID \in \{\text{TRUE}, \text{FALSE}\}$
Creating sign-by-sample matrices
$\omega^{(\text{s})} = 0.5, \ \omega^{(\text{v})} = 0.5$
Sign-by-sample matrices

Figure S1. Detailed workflow of Figure 1c focusing on the parameter settings. The indicated values are preset as default in ASURAT, while "u.d." stands for the value or argument that users must define. Here, α and β are positive and negative threshold values of correlation coefficients; n_{min} and $n^{(\text{w})}_{\text{min}}$, positive integers for selecting reliable signs; *MEASURE*, the name of information content (IC)-based method defining semantic similarities; *SIM_TH*, a threshold value used to regard two biological terms as similar; *KEEP_RAREID* determines whether the signs with larger ICs are kept or not (if TRUE, the signs with larger ICs are kept), and $\omega^{(\text{s})}$ and $\omega^{(\text{v})}$ weight constants are used to define sign-by-sample matrices.

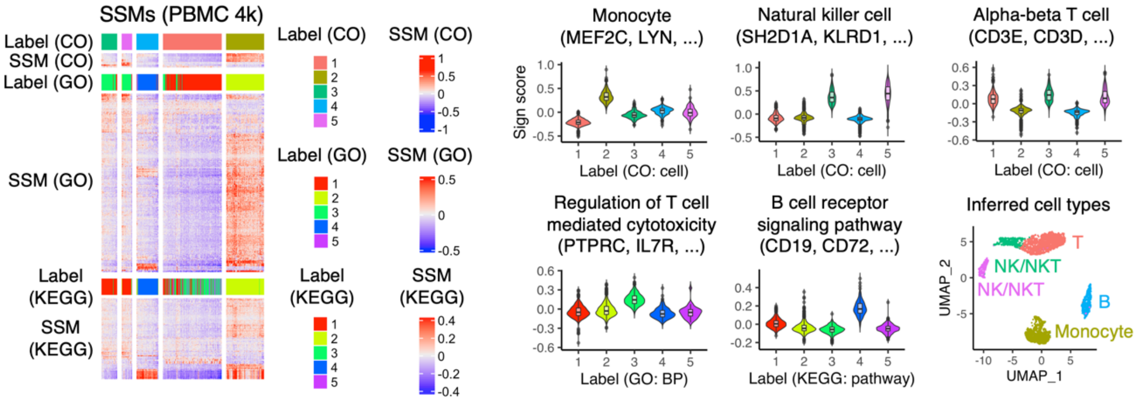


Figure S2. Clustering peripheral blood mononuclear cell (PBMC) 4k single-cell transcriptomes using ASURAT. Identification of cell types in the PBMC 4k dataset from analyses of sign-by-sample matrices (SSMs) for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG). According to heatmaps and violin plots, showing representative signs and the functional gene sets, T cell ("T"), B cell ("B"), NK or NKT cell ("NK/NKT"), and monocyte were identified as shown in Uniform Manifold Approximation and Projection (UMAP) plots.

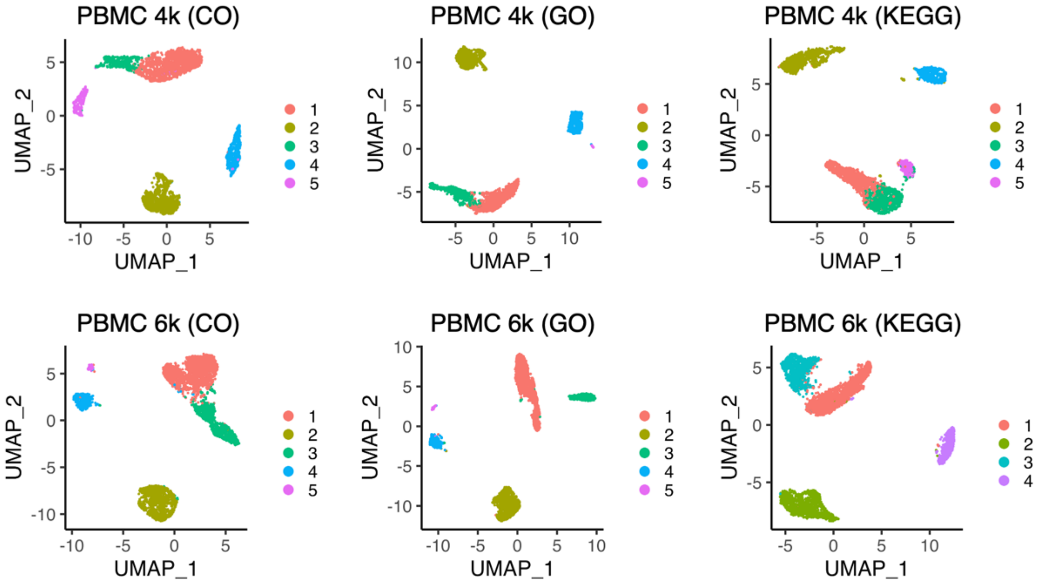


Figure S3. Clustering peripheral blood mononuclear cell (PBMC) 4k and 6k single-cell transcriptomes using ASURAT. Uniform Manifold Approximation and Projection (UMAP) plots of sign-by-sample matrices for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG).

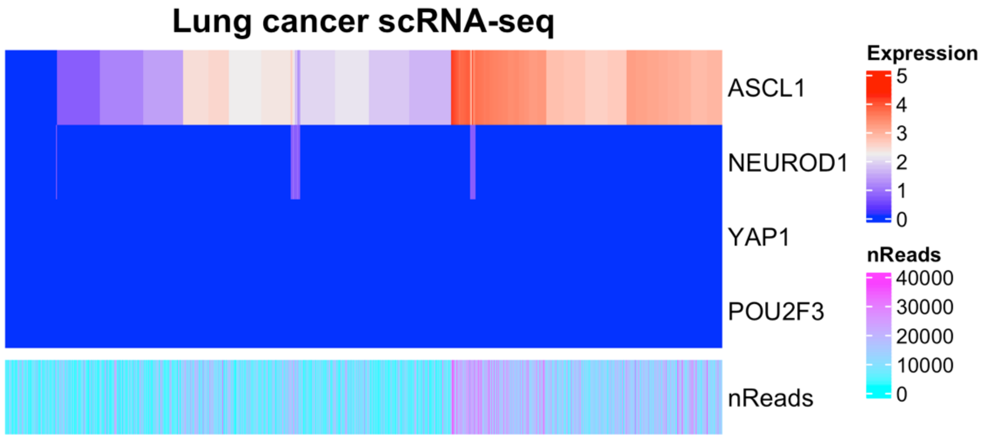


Figure S4. Heatmaps of expression levels of known small cell lung cancer marker genes. Log-normalized gene expression levels for *ASCL1*, *NEUROD1*, *YAP1*, and *POU2F3* across all the cells after controlling for data quality. nReads, total read counts.

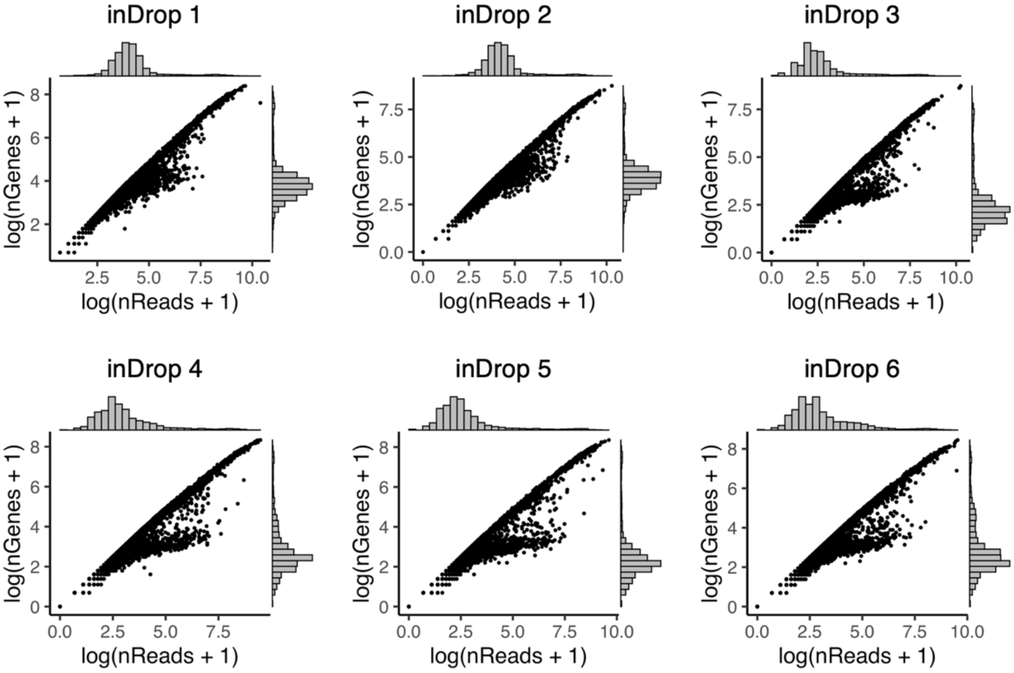
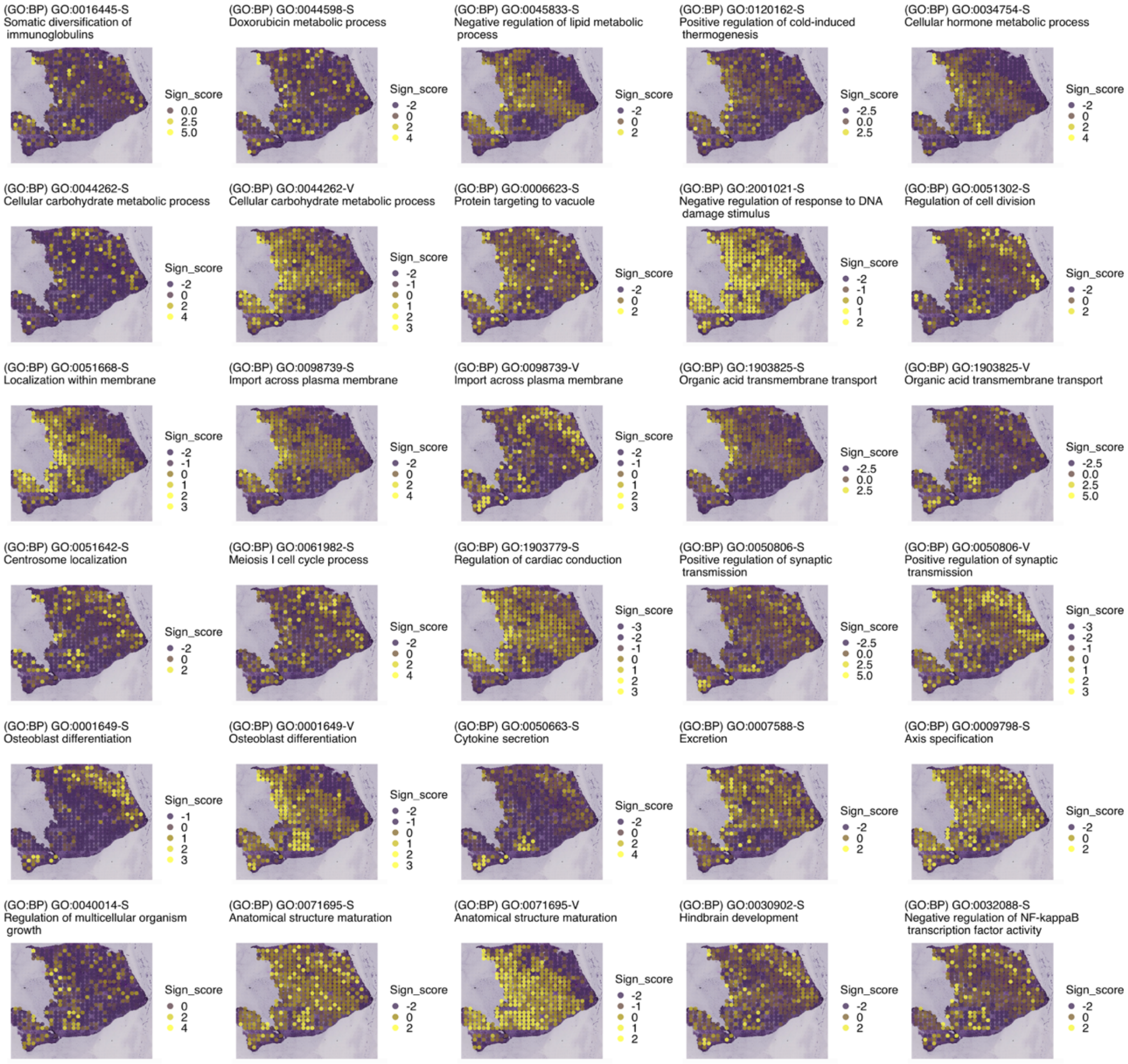
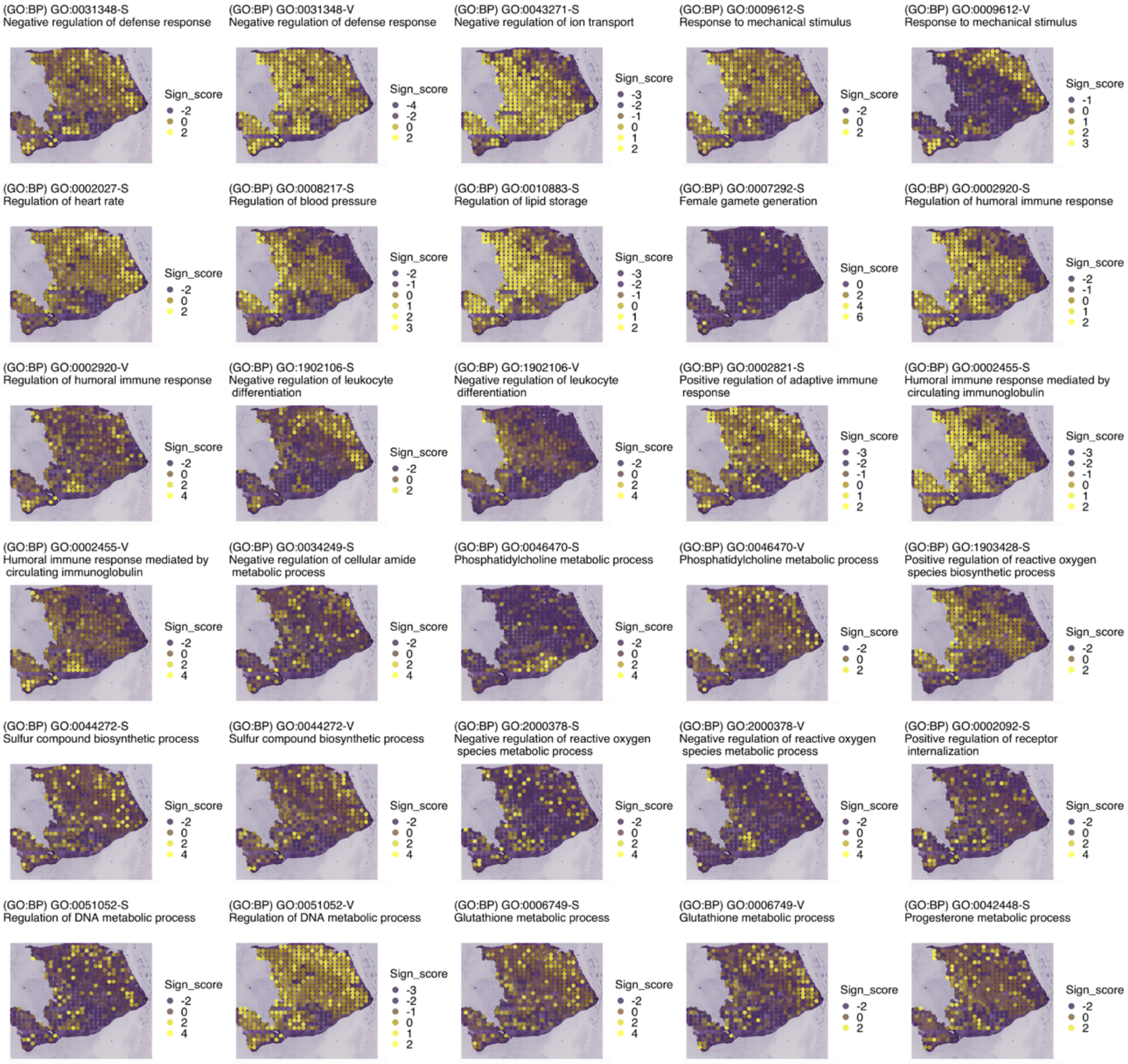
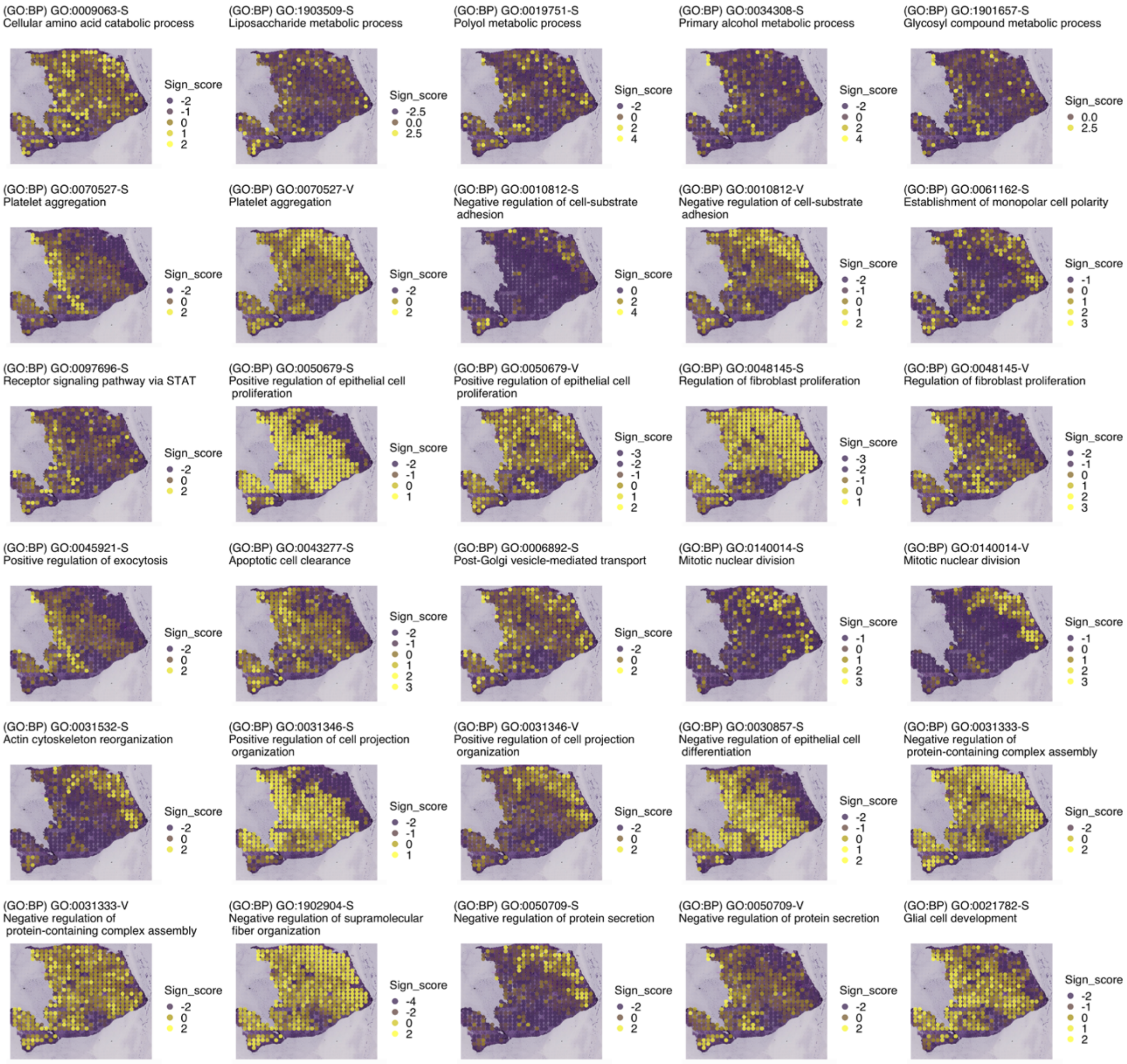
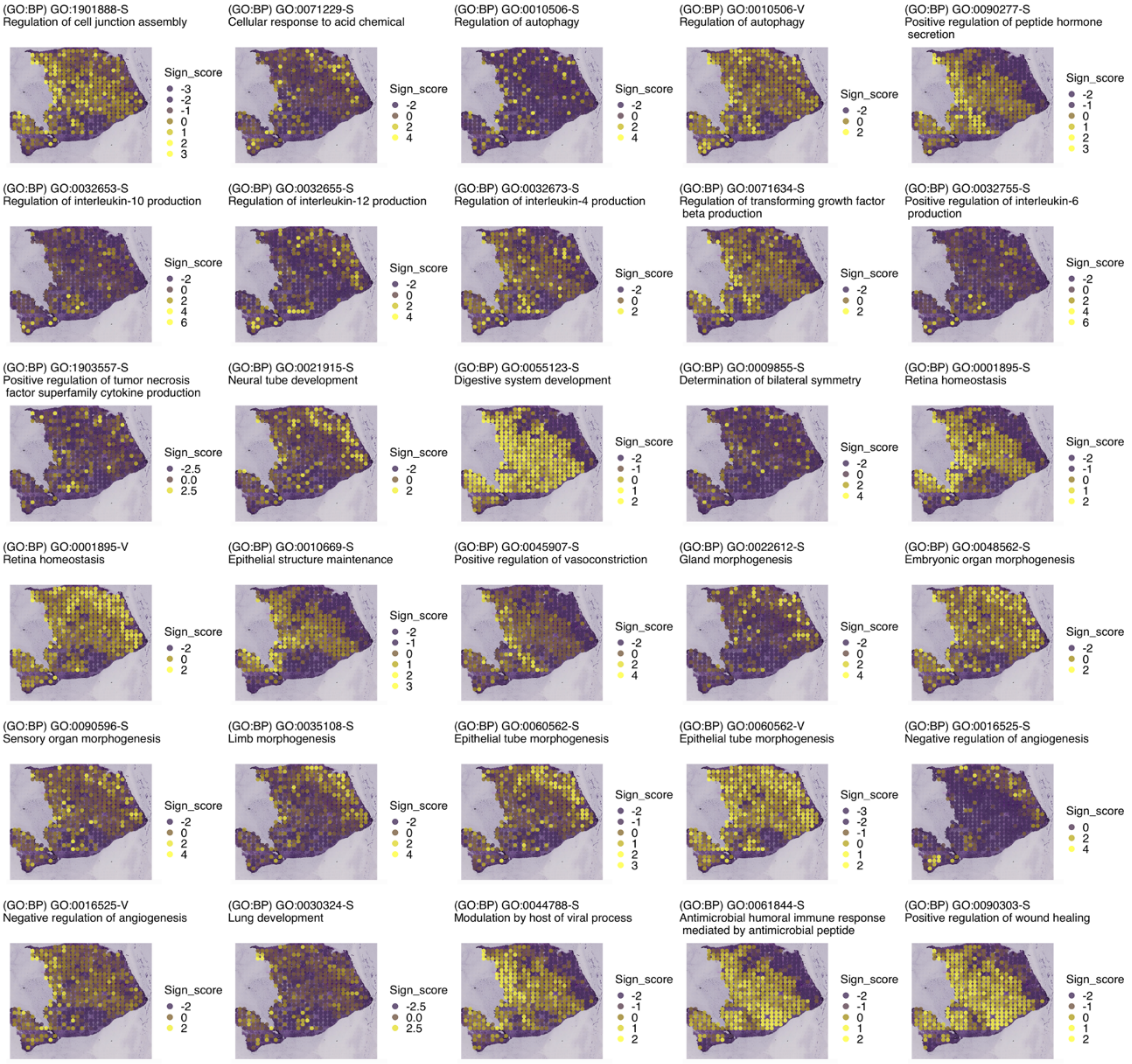


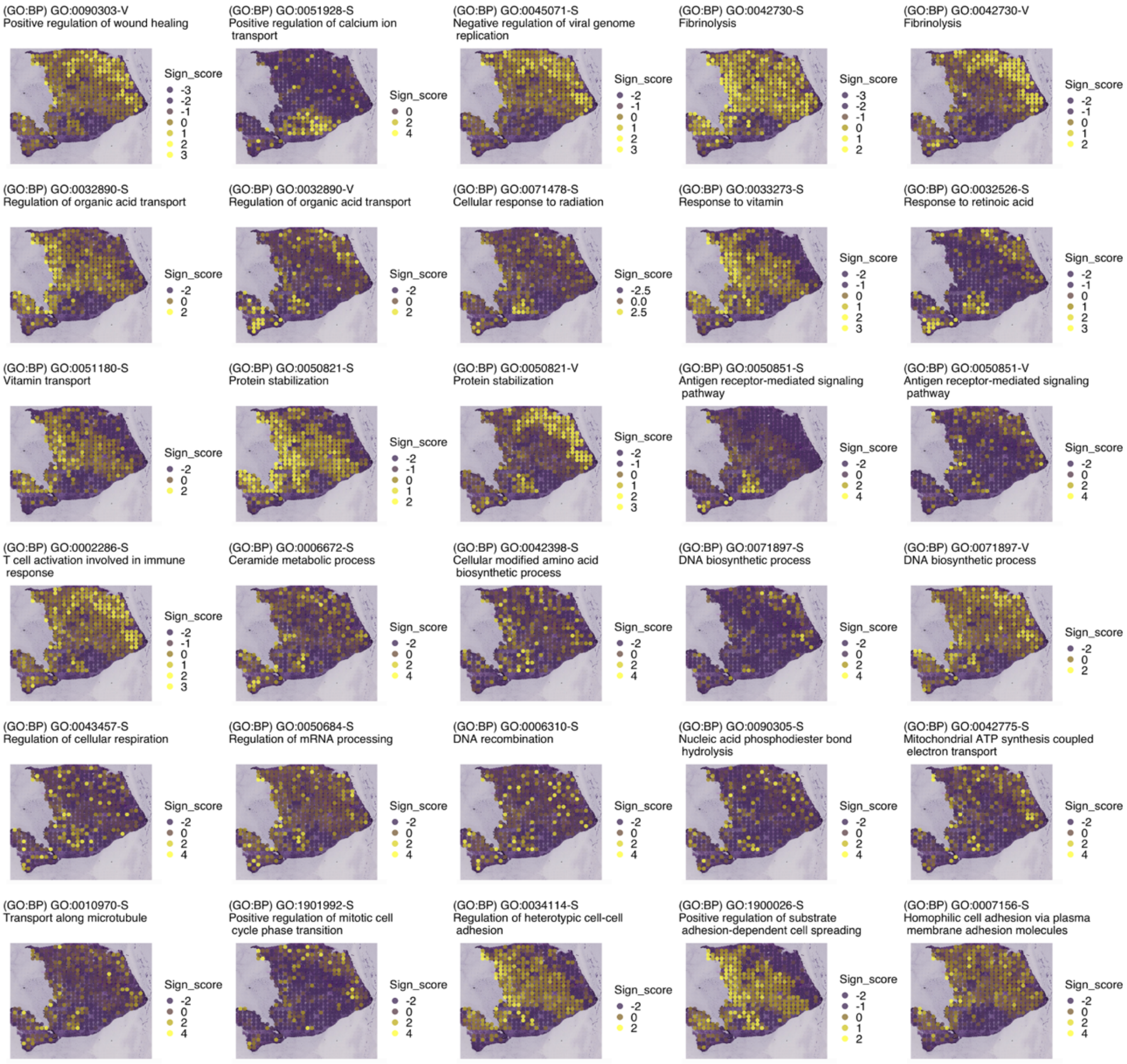
Figure S5. Data qualities across all the cells in single-cell RNA sequencing datasets PDAC-A inDrop from 1 to 6. nReads, total read counts; nGenes, number of genes expressed with non-zero read counts.

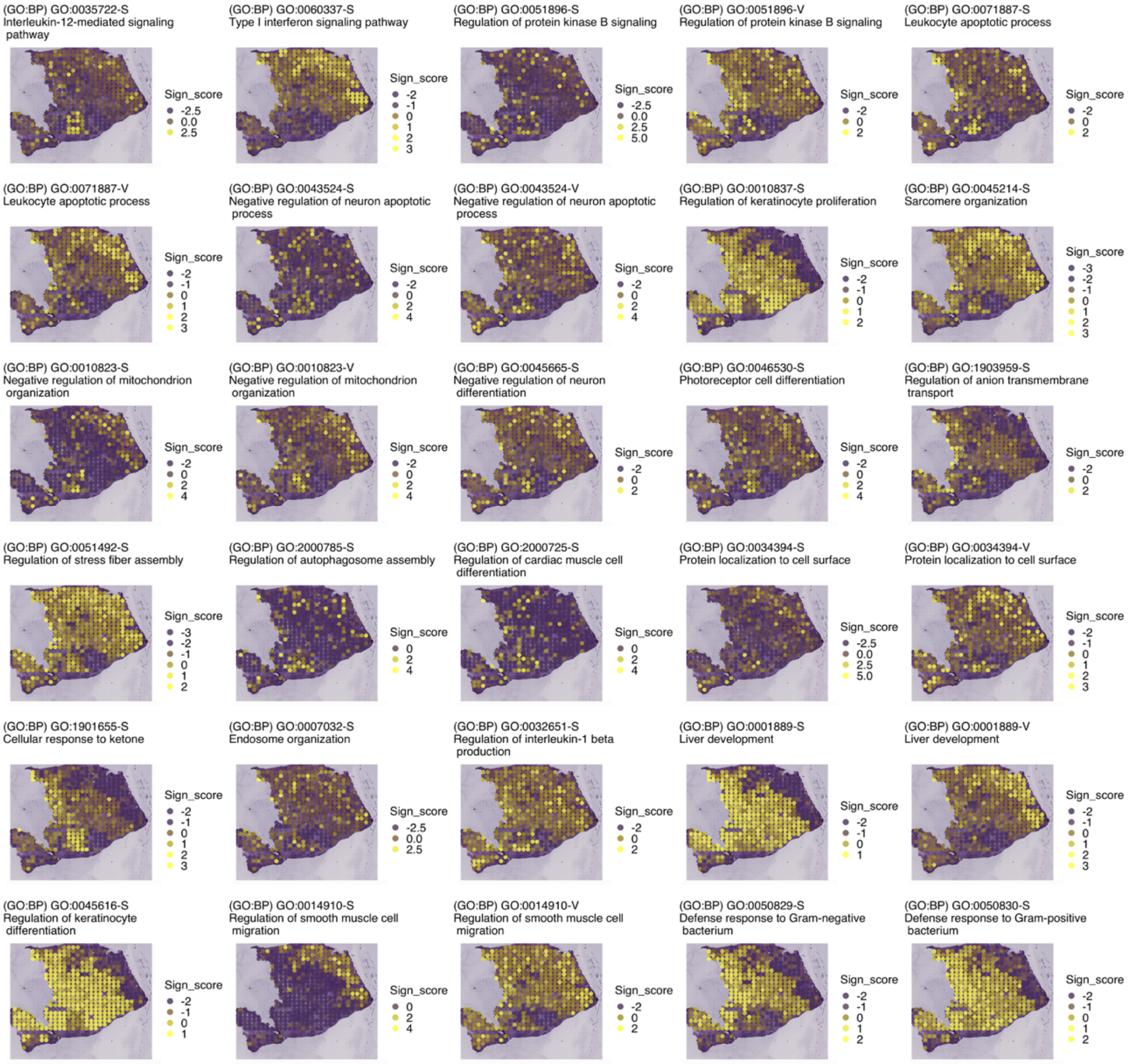

(GO:BP) GO:0016445-S Somatic diversification of immunoglobulins
Sign_score 0.0 2.5 5.0
(GO:BP) GO:0044598-S Doxorubicin metabolic process
Sign_score -2 0 2 4
(GO:BP) GO:0045833-S Negative regulation of lipid metabolic process
Sign_score -2 0 2
(GO:BP) GO:0120162-S Positive regulation of cold-induced thermogenesis
Sign_score -2.5 0.0 2.5
(GO:BP) GO:0034754-S Cellular hormone metabolic process
Sign_score -2 0 2 4
(GO:BP) GO:0044262-S Cellular carbohydrate metabolic process
Sign_score -2 0 2 4
(GO:BP) GO:0044262-V Cellular carbohydrate metabolic process
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0006623-S Protein targeting to vacuole
Sign_score -2 0 2
(GO:BP) GO:2001021-S Negative regulation of response to DNA damage stimulus
Sign_score -2 -1 0 1 2
(GO:BP) GO:0051302-S Regulation of cell division
Sign_score -2 0 2
(GO:BP) GO:0051668-S Localization within membrane
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0098739-S Import across plasma membrane
Sign_score -2 0 2 4
(GO:BP) GO:0098739-V Import across plasma membrane
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:1903825-S Organic acid transmembrane transport
Sign_score -2.5 0.0 2.5
(GO:BP) GO:1903825-V Organic acid transmembrane transport
Sign_score -2.5 0.0 2.5 5.0
(GO:BP) GO:0051642-S Centrosome localization
Sign_score -2 0 2
(GO:BP) GO:0061982-S Meiosis I cell cycle process
Sign_score -2 0 2 4
(GO:BP) GO:1903779-S Regulation of cardiac conduction
Sign_score -3 -2 -1 0 1 2 3
(GO:BP) GO:0050806-S Positive regulation of synaptic transmission
Sign_score -2.5 0.0 2.5 5.0
(GO:BP) GO:0050806-V Positive regulation of synaptic transmission
Sign_score -3 -2 -1 0 1 2 3
(GO:BP) GO:0001649-S Osteoblast differentiation
Sign_score -1 0 1 2 3
(GO:BP) GO:0001649-V Osteoblast differentiation
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0050663-S Cytokine secretion
Sign_score -2 0 2 4
(GO:BP) GO:0007588-S Excretion
Sign_score -2 0 2
(GO:BP) GO:0009798-S Axis specification
Sign_score -2 0 2
(GO:BP) GO:0040014-S Regulation of multicellular organism growth
Sign_score 0 2 4
(GO:BP) GO:0071695-S Anatomical structure maturation
Sign_score -2 0 2
(GO:BP) GO:0071695-V Anatomical structure maturation
Sign_score -2 -1 0 1 2
(GO:BP) GO:0030902-S Hindbrain development
Sign_score -2 0 2
(GO:BP) GO:0032088-S Negative regulation of NF-kappaB transcription factor activity
Sign_score -2 0 2

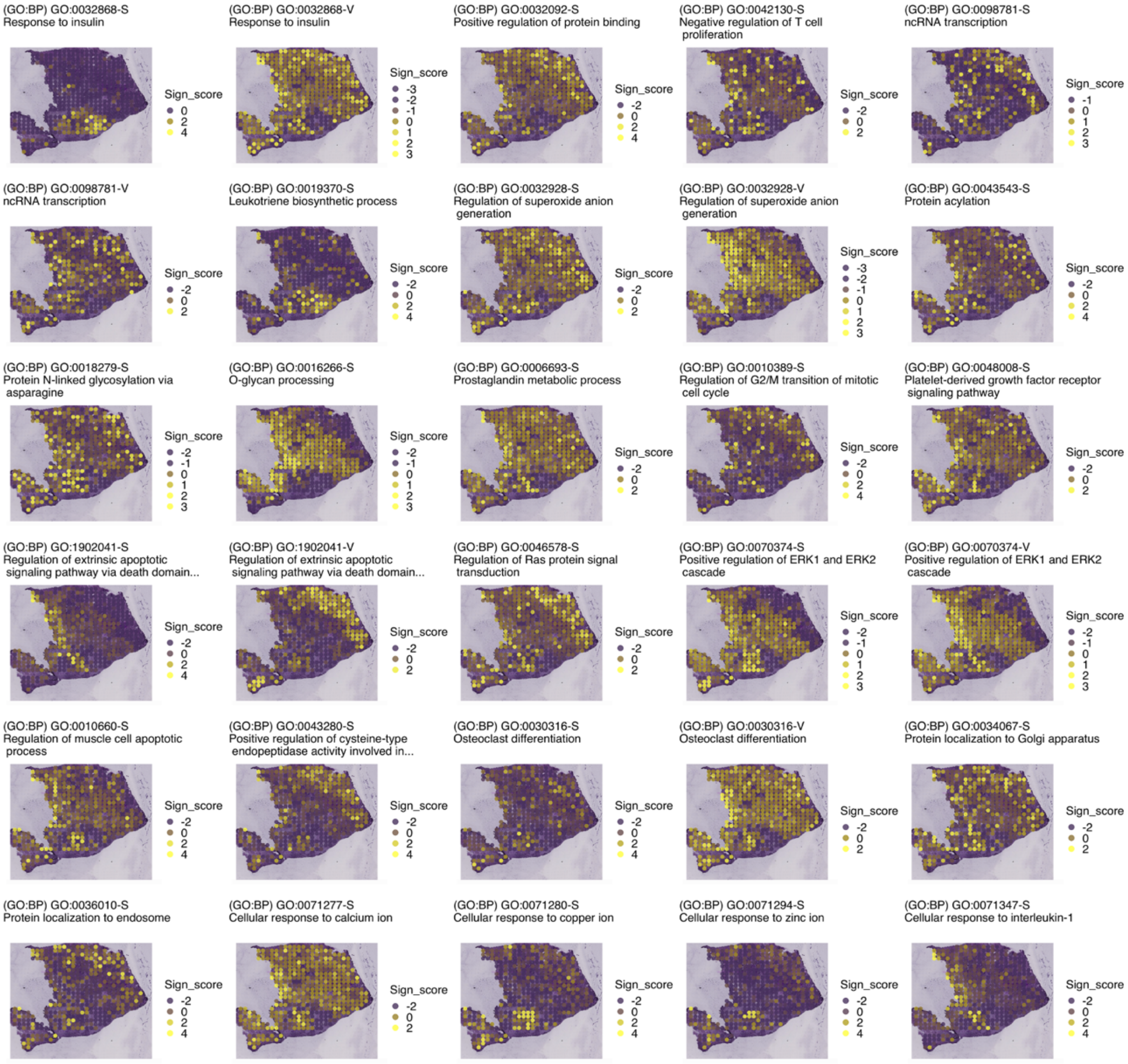

(GO:BP) GO:0031348-S
Negative regulation of defense response
Sign_score
-2
0
2
(GO:BP) GO:0031348-V
Negative regulation of defense response
Sign_score
-4
-2
0
2
(GO:BP) GO:0043271-S
Negative regulation of ion transport
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0009612-S
Response to mechanical stimulus
Sign_score
-2
0
2
(GO:BP) GO:0009612-V
Response to mechanical stimulus
Sign_score
-1
0
1
2
3
(GO:BP) GO:0002027-S
Regulation of heart rate
Sign_score
-2
0
2
(GO:BP) GO:0008217-S
Regulation of blood pressure
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0010883-S
Regulation of lipid storage
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0007292-S
Female gamete generation
Sign_score
0
2
4
6
(GO:BP) GO:0002920-S
Regulation of humoral immune response
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0002920-V
Regulation of humoral immune response
Sign_score
-2
0
2
4
(GO:BP) GO:1902106-S
Negative regulation of leukocyte differentiation
Sign_score
-2
0
2
(GO:BP) GO:1902106-V
Negative regulation of leukocyte differentiation
Sign_score
-2
0
2
4
(GO:BP) GO:0002821-S
Positive regulation of adaptive immune response
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0002455-S
Humoral immune response mediated by circulating immunoglobulin
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0002455-V
Humoral immune response mediated by circulating immunoglobulin
Sign_score
-2
0
2
4
(GO:BP) GO:0034249-S
Negative regulation of cellular amide metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:0046470-S
Phosphatidylcholine metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:0046470-V
Phosphatidylcholine metabolic process
Sign_score
-2
0
2
(GO:BP) GO:1903428-S
Positive regulation of reactive oxygen species biosynthetic process
Sign_score
-2
0
2
(GO:BP) GO:0044272-S
Sulfur compound biosynthetic process
Sign_score
-2
0
2
4
(GO:BP) GO:0044272-V
Sulfur compound biosynthetic process
Sign_score
-2
0
2
4
(GO:BP) GO:2000378-S
Negative regulation of reactive oxygen species metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:2000378-V
Negative regulation of reactive oxygen species metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:0002092-S
Positive regulation of receptor internalization
Sign_score
-2
0
2
4
(GO:BP) GO:0051052-S
Regulation of DNA metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:0051052-V
Regulation of DNA metabolic process
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0006749-S
Glutathione metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:0006749-V
Glutathione metabolic process
Sign_score
-2
0
2
(GO:BP) GO:0042448-S
Progesterone metabolic process
Sign_score
-2
0
2


(GO:BP) GO:0009063-S
Cellular amino acid catabolic process
Sign_score
-2
-1
0
1
2
(GO:BP) GO:1903509-S
Liposaccharide metabolic process
Sign_score
-2.5
0.0
2.5
(GO:BP) GO:0019751-S
Polyol metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:0034308-S
Primary alcohol metabolic process
Sign_score
-2
0
2
4
(GO:BP) GO:1901657-S
Glycosyl compound metabolic process
Sign_score
0.0
2.5
(GO:BP) GO:0070527-S
Platelet aggregation
Sign_score
-2
0
2
(GO:BP) GO:0070527-V
Platelet aggregation
Sign_score
-2
0
2
(GO:BP) GO:0010812-S
Negative regulation of cell-substrate adhesion
Sign_score
0
2
4
(GO:BP) GO:0010812-V
Negative regulation of cell-substrate adhesion
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0061162-S
Establishment of monopolar cell polarity
Sign_score
-1
0
1
2
3
(GO:BP) GO:0097696-S
Receptor signaling pathway via STAT
Sign_score
-2
0
2
(GO:BP) GO:0050679-S
Positive regulation of epithelial cell proliferation
Sign_score
-2
-1
0
1
(GO:BP) GO:0050679-V
Positive regulation of epithelial cell proliferation
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0048145-S
Regulation of fibroblast proliferation
Sign_score
-3
-2
-1
0
1
(GO:BP) GO:0048145-V
Regulation of fibroblast proliferation
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0045921-S
Positive regulation of exocytosis
Sign_score
-2
0
2
(GO:BP) GO:0043277-S
Apoptotic cell clearance
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0006892-S
Post-Golgi vesicle-mediated transport
Sign_score
-2
0
2
(GO:BP) GO:0140014-S
Mitotic nuclear division
Sign_score
-1
0
1
2
3
(GO:BP) GO:0140014-V
Mitotic nuclear division
Sign_score
-1
0
1
2
3
(GO:BP) GO:0031532-S
Actin cytoskeleton reorganization
Sign_score
-2
0
2
(GO:BP) GO:0031346-S
Positive regulation of cell projection organization
Sign_score
-2
-1
0
1
(GO:BP) GO:0031346-V
Positive regulation of cell projection organization
Sign_score
-2
0
2
(GO:BP) GO:0030857-S
Negative regulation of epithelial cell differentiation
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0031333-S
Negative regulation of protein-containing complex assembly
Sign_score
-2
0
2
(GO:BP) GO:0031333-V
Negative regulation of protein-containing complex assembly
Sign_score
-2
0
2
(GO:BP) GO:1902904-S
Negative regulation of supramolecular fiber organization
Sign_score
-4
-2
0
2
(GO:BP) GO:0050709-S
Negative regulation of protein secretion
Sign_score
-2
0
2
(GO:BP) GO:0050709-V
Negative regulation of protein secretion
Sign_score
-2
0
2
(GO:BP) GO:0021782-S
Glial cell development
Sign_score
-2
-1
0
1
2


(GO:BP) GO:1901888-S
Regulation of cell junction assembly
Sign_score
-3
-2
-1
0
1
2
3
(GO:BP) GO:0071229-S
Cellular response to acid chemical
Sign_score
-2
0
2
4
(GO:BP) GO:0010506-S
Regulation of autophagy
Sign_score
-2
0
2
4
(GO:BP) GO:0010506-V
Regulation of autophagy
Sign_score
-2
0
2
(GO:BP) GO:0090277-S
Positive regulation of peptide hormone secretion
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0032653-S
Regulation of interleukin-10 production
Sign_score
-2
0
2
4
6
(GO:BP) GO:0032655-S
Regulation of interleukin-12 production
Sign_score
-2
0
2
4
(GO:BP) GO:0032673-S
Regulation of interleukin-4 production
Sign_score
-2
0
2
(GO:BP) GO:0071634-S
Regulation of transforming growth factor beta production
Sign_score
-2
0
2
(GO:BP) GO:0032755-S
Positive regulation of interleukin-6 production
Sign_score
-2
0
2
4
6
(GO:BP) GO:1903557-S
Positive regulation of tumor necrosis factor superfamily cytokine production
Sign_score
-2.5
0.0
2.5
(GO:BP) GO:0021915-S
Neural tube development
Sign_score
-2
0
2
(GO:BP) GO:0055123-S
Digestive system development
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0009855-S
Determination of bilateral symmetry
Sign_score
-2
0
2
4
(GO:BP) GO:0001895-S
Retina homeostasis
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0001895-V
Retina homeostasis
Sign_score
-2
0
2
(GO:BP) GO:0010669-S
Epithelial structure maintenance
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0045907-S
Positive regulation of vasoconstriction
Sign_score
-2
0
2
4
(GO:BP) GO:0022612-S
Gland morphogenesis
Sign_score
-2
0
2
4
(GO:BP) GO:0048562-S
Embryonic organ morphogenesis
Sign_score
-2
0
2
(GO:BP) GO:0090596-S
Sensory organ morphogenesis
Sign_score
-2
0
2
4
(GO:BP) GO:0035108-S
Limb morphogenesis
Sign_score
-2
0
2
4
(GO:BP) GO:0060562-S
Epithelial tube morphogenesis
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0060562-V
Epithelial tube morphogenesis
Sign_score
-3
-2
-1
0
1
2
(GO:BP) GO:0016525-S
Negative regulation of angiogenesis
Sign_score
0
2
4
(GO:BP) GO:0016525-V
Negative regulation of angiogenesis
Sign_score
-2
0
2
(GO:BP) GO:0030324-S
Lung development
Sign_score
-2.5
0.0
2.5
(GO:BP) GO:0044788-S
Modulation by host of viral process
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0061844-S
Antimicrobial humoral immune response mediated by antimicrobial peptide
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0090303-S
Positive regulation of wound healing
Sign_score
-2
-1
0
1
2


(GO:BP) GO:0090303-V
Positive regulation of wound healing
Sign_score -3 -2 -1 0 1 2 3
(GO:BP) GO:0051928-S
Positive regulation of calcium ion transport
Sign_score 0 2 4
(GO:BP) GO:0045071-S
Negative regulation of viral genome replication
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0042730-S
Fibrinolysis
Sign_score -3 -2 -1 0 1 2
(GO:BP) GO:0042730-V
Fibrinolysis
Sign_score -2 -1 0 1 2
(GO:BP) GO:0032890-S
Regulation of organic acid transport
Sign_score -2 0 2
(GO:BP) GO:0032890-V
Regulation of organic acid transport
Sign_score -2 0 2
(GO:BP) GO:0071478-S
Cellular response to radiation
Sign_score -2.5 0.0 2.5
(GO:BP) GO:0033273-S
Response to vitamin
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0032526-S
Response to retinoic acid
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0051180-S
Vitamin transport
Sign_score -2 0 2
(GO:BP) GO:0050821-S
Protein stabilization
Sign_score -2 -1 0 1 2
(GO:BP) GO:0050821-V
Protein stabilization
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0050851-S
Antigen receptor-mediated signaling pathway
Sign_score -2 0 2 4
(GO:BP) GO:0050851-V
Antigen receptor-mediated signaling pathway
Sign_score -2 0 2 4
(GO:BP) GO:0002286-S
T cell activation involved in immune response
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0006672-S
Ceramide metabolic process
Sign_score -2 0 2 4
(GO:BP) GO:0042398-S
Cellular modified amino acid biosynthetic process
Sign_score -2 0 2 4
(GO:BP) GO:0071897-S
DNA biosynthetic process
Sign_score -2 0 2 4
(GO:BP) GO:0071897-V
DNA biosynthetic process
Sign_score -2 0 2
(GO:BP) GO:0043457-S
Regulation of cellular respiration
Sign_score -2 0 2 4
(GO:BP) GO:0050684-S
Regulation of mRNA processing
Sign_score -2 0 2 4
(GO:BP) GO:0006310-S
DNA recombination
Sign_score -2 0 2 4
(GO:BP) GO:0090305-S
Nucleic acid phosphodiester bond hydrolysis
Sign_score -2 0 2 4
(GO:BP) GO:0042775-S
Mitochondrial ATP synthesis coupled electron transport
Sign_score -2 0 2 4
(GO:BP) GO:0010970-S
Transport along microtubule
Sign_score -2 0 2 4
(GO:BP) GO:1901992-S
Positive regulation of mitotic cell cycle phase transition
Sign_score -2 0 2 4
(GO:BP) GO:0034114-S
Regulation of heterotypic cell-cell adhesion
Sign_score -2 0 2
(GO:BP) GO:1900026-S
Positive regulation of substrate adhesion-dependent cell spreading
Sign_score -2 -1 0 1 2
(GO:BP) GO:0007156-S
Homophilic cell adhesion via plasma membrane adhesion molecules
Sign_score -2 0 2 4


(GO:BP) GO:0035722-S Interleukin-12-mediated signaling pathway
Sign_score -2.5 0.0 2.5
(GO:BP) GO:0060337-S Type I interferon signaling pathway
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0051896-S Regulation of protein kinase B signaling
Sign_score -2.5 0.0 2.5 5.0
(GO:BP) GO:0051896-V Regulation of protein kinase B signaling
Sign_score -2 0 2
(GO:BP) GO:0071887-S Leukocyte apoptotic process
Sign_score -2 0 2
(GO:BP) GO:0071887-V Leukocyte apoptotic process
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0043524-S Negative regulation of neuron apoptotic process
Sign_score -2 0 2 4
(GO:BP) GO:0043524-V Negative regulation of neuron apoptotic process
Sign_score -2 0 2 4
(GO:BP) GO:0010837-S Regulation of keratinocyte proliferation
Sign_score -2 -1 0 1 2
(GO:BP) GO:0045214-S Sarcomere organization
Sign_score -3 -2 -1 0 1 2 3
(GO:BP) GO:0010823-S Negative regulation of mitochondrion organization
Sign_score -2 0 2 4
(GO:BP) GO:0010823-V Negative regulation of mitochondrion organization
Sign_score -2 0 2 4
(GO:BP) GO:0045665-S Negative regulation of neuron differentiation
Sign_score -2 0 2
(GO:BP) GO:0046530-S Photoreceptor cell differentiation
Sign_score -2 0 2 4
(GO:BP) GO:1903959-S Regulation of anion transmembrane transport
Sign_score -2 0 2
(GO:BP) GO:0051492-S Regulation of stress fiber assembly
Sign_score -3 -2 -1 0 1 2
(GO:BP) GO:2000785-S Regulation of autophagosome assembly
Sign_score 0 2 4
(GO:BP) GO:2000725-S Regulation of cardiac muscle cell differentiation
Sign_score 0 2 4
(GO:BP) GO:0034394-S Protein localization to cell surface
Sign_score -2.5 0.0 2.5 5.0
(GO:BP) GO:0034394-V Protein localization to cell surface
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:1901655-S Cellular response to ketone
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0007032-S Endosome organization
Sign_score -2.5 0.0 2.5
(GO:BP) GO:0032651-S Regulation of interleukin-1 beta production
Sign_score -2 0 2
(GO:BP) GO:0001889-S Liver development
Sign_score -2 -1 0 1
(GO:BP) GO:0001889-V Liver development
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0045616-S Regulation of keratinocyte differentiation
Sign_score -2 -1 0 1
(GO:BP) GO:0014910-S Regulation of smooth muscle cell migration
Sign_score 0 2 4
(GO:BP) GO:0014910-V Regulation of smooth muscle cell migration
Sign_score -2 0 2
(GO:BP) GO:0050829-S Defense response to Gram-negative bacterium
Sign_score -2 -1 0 1 2
(GO:BP) GO:0050830-S Defense response to Gram-positive bacterium
Sign_score -2 -1 0 1 2



(GO:BP) GO:0032868-S
Response to insulin

(GO:BP) GO:0032868-V
Response to insulin

(GO:BP) GO:0032092-S
Positive regulation of protein binding

(GO:BP) GO:0042130-S
Negative regulation of T cell proliferation

(GO:BP) GO:0098781-S
ncRNA transcription

(GO:BP) GO:0098781-V
ncRNA transcription

(GO:BP) GO:0019370-S
Leukotriene biosynthetic process

(GO:BP) GO:0032928-S
Regulation of superoxide anion generation

(GO:BP) GO:0032928-V
Regulation of superoxide anion generation

(GO:BP) GO:0043543-S
Protein acylation

(GO:BP) GO:0018279-S
Protein N-linked glycosylation via asparagine

(GO:BP) GO:0016266-S
O-glycan processing

(GO:BP) GO:0006693-S
Prostaglandin metabolic process

(GO:BP) GO:0010389-S
Regulation of G2/M transition of mitotic cell cycle

(GO:BP) GO:0048008-S
Platelet-derived growth factor receptor signaling pathway

(GO:BP) GO:1902041-S
Regulation of extrinsic apoptotic signaling pathway via death domain...

(GO:BP) GO:1902041-V
Regulation of extrinsic apoptotic signaling pathway via death domain...

(GO:BP) GO:0046578-S
Regulation of Ras protein signal transduction

(GO:BP) GO:0070374-S
Positive regulation of ERK1 and ERK2 cascade

(GO:BP) GO:0070374-V
Positive regulation of ERK1 and ERK2 cascade

(GO:BP) GO:0010660-S
Regulation of muscle cell apoptotic process

(GO:BP) GO:0043280-S
Positive regulation of cysteine-type endopeptidase activity involved in...

(GO:BP) GO:0030316-S
Osteoclast differentiation

(GO:BP) GO:0030316-V
Osteoclast differentiation

(GO:BP) GO:0034067-S
Protein localization to Golgi apparatus

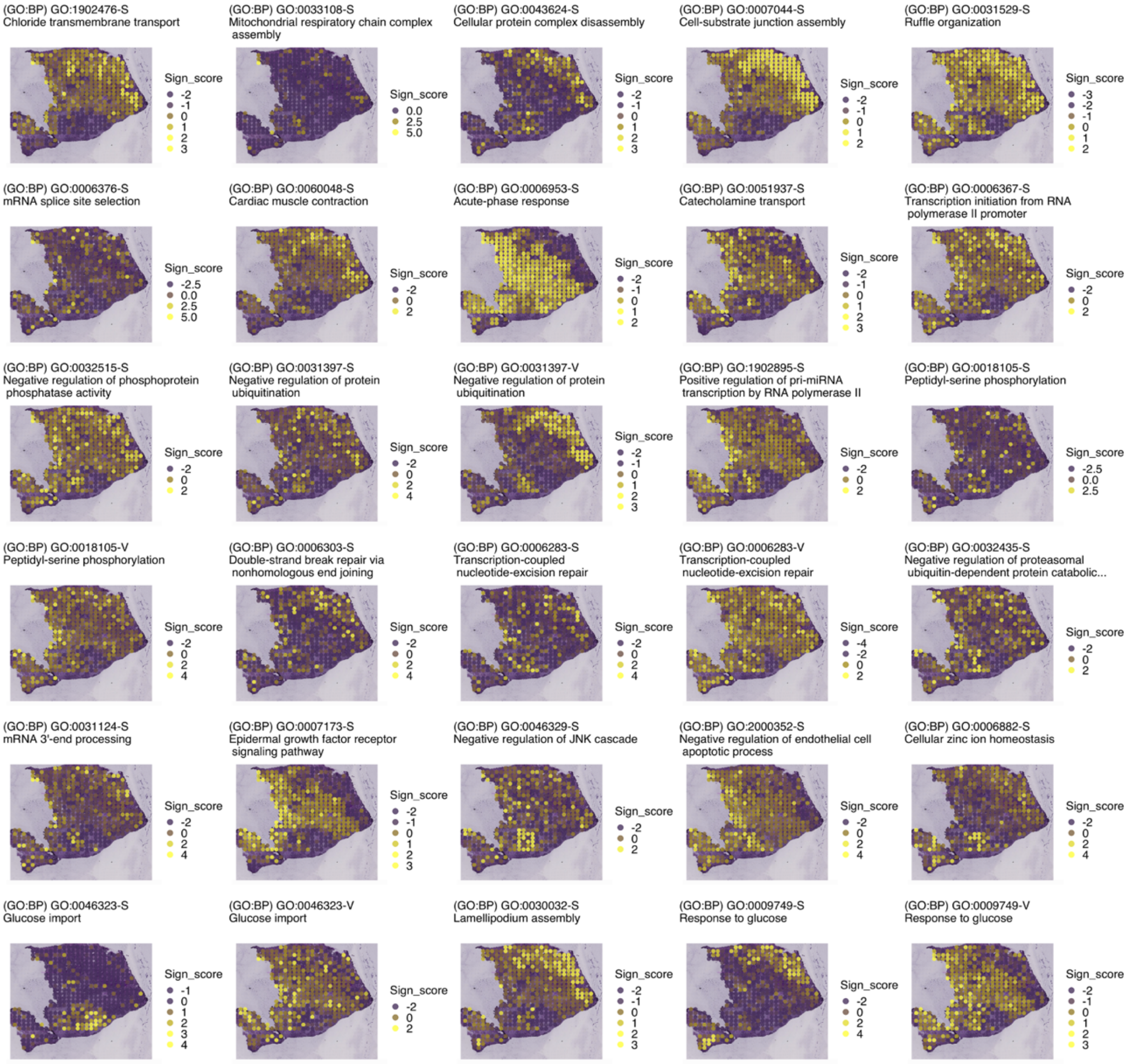
(GO:BP) GO:0036010-S
Protein localization to endosome

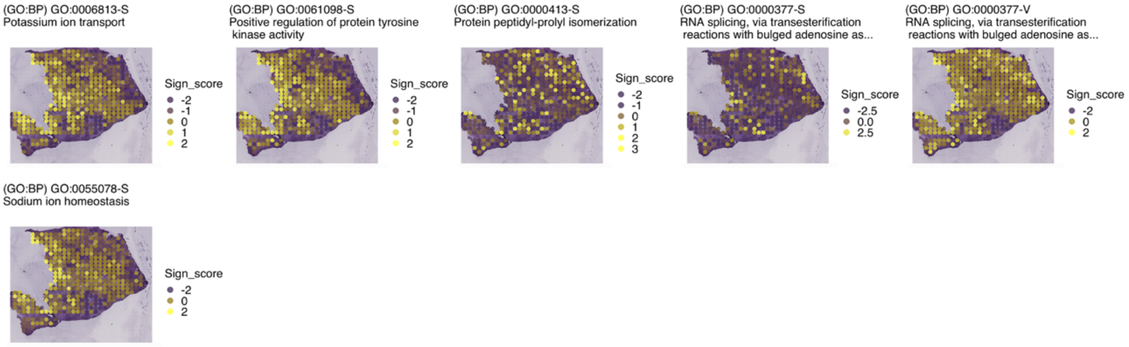
(GO:BP) GO:0071277-S
Cellular response to calcium ion

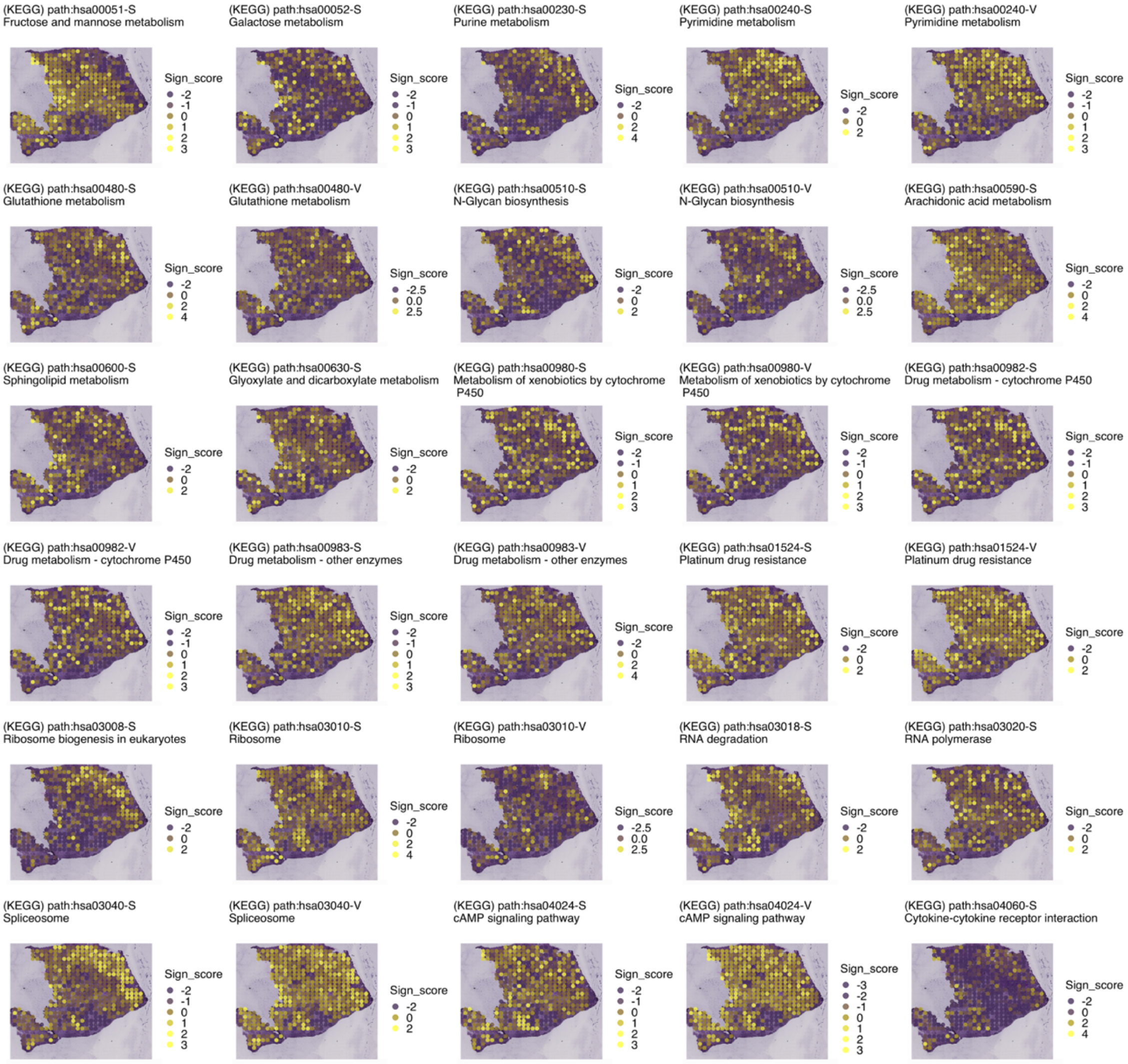
(GO:BP) GO:0071280-S
Cellular response to copper ion

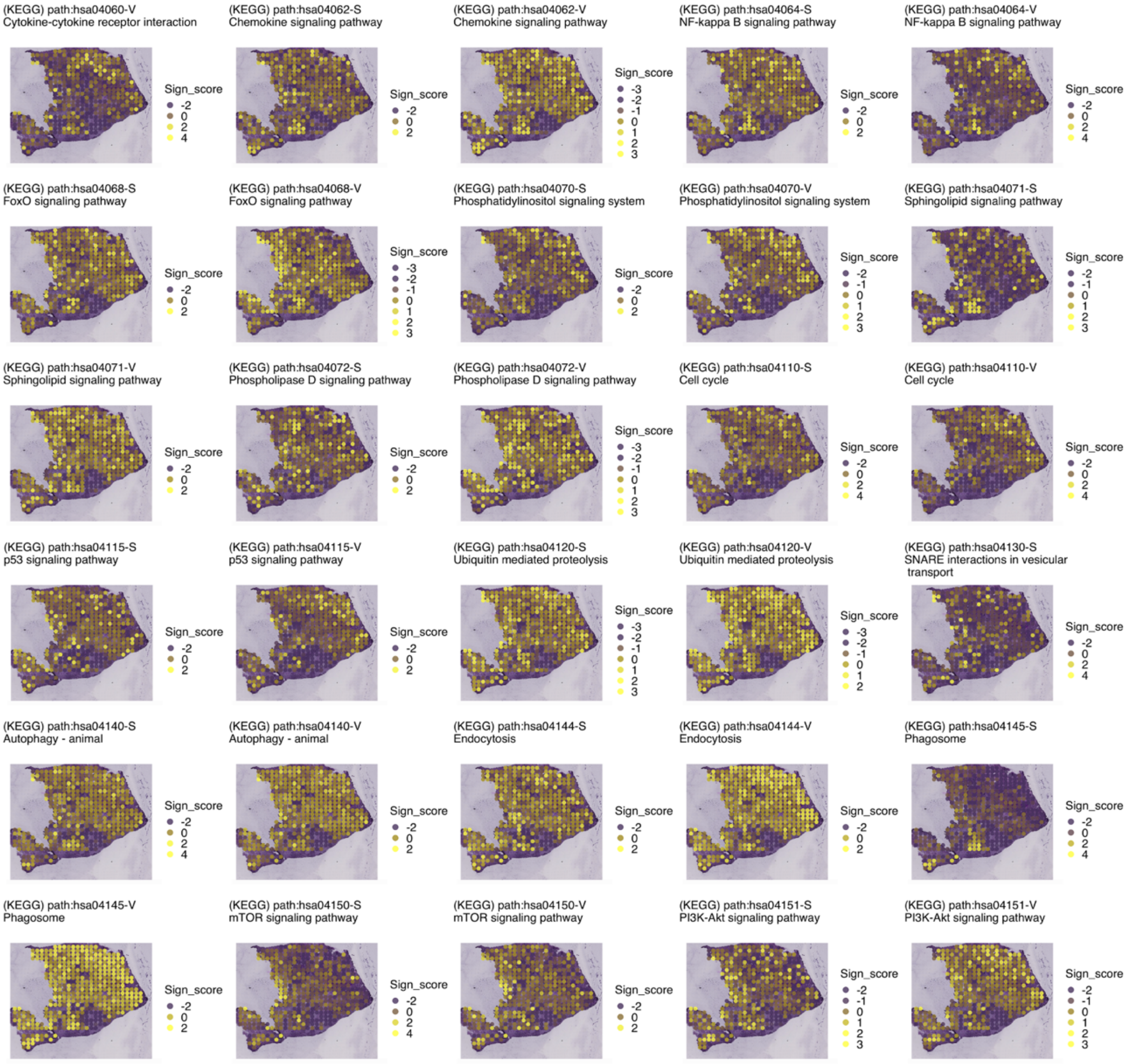
(GO:BP) GO:0071294-S
Cellular response to zinc ion

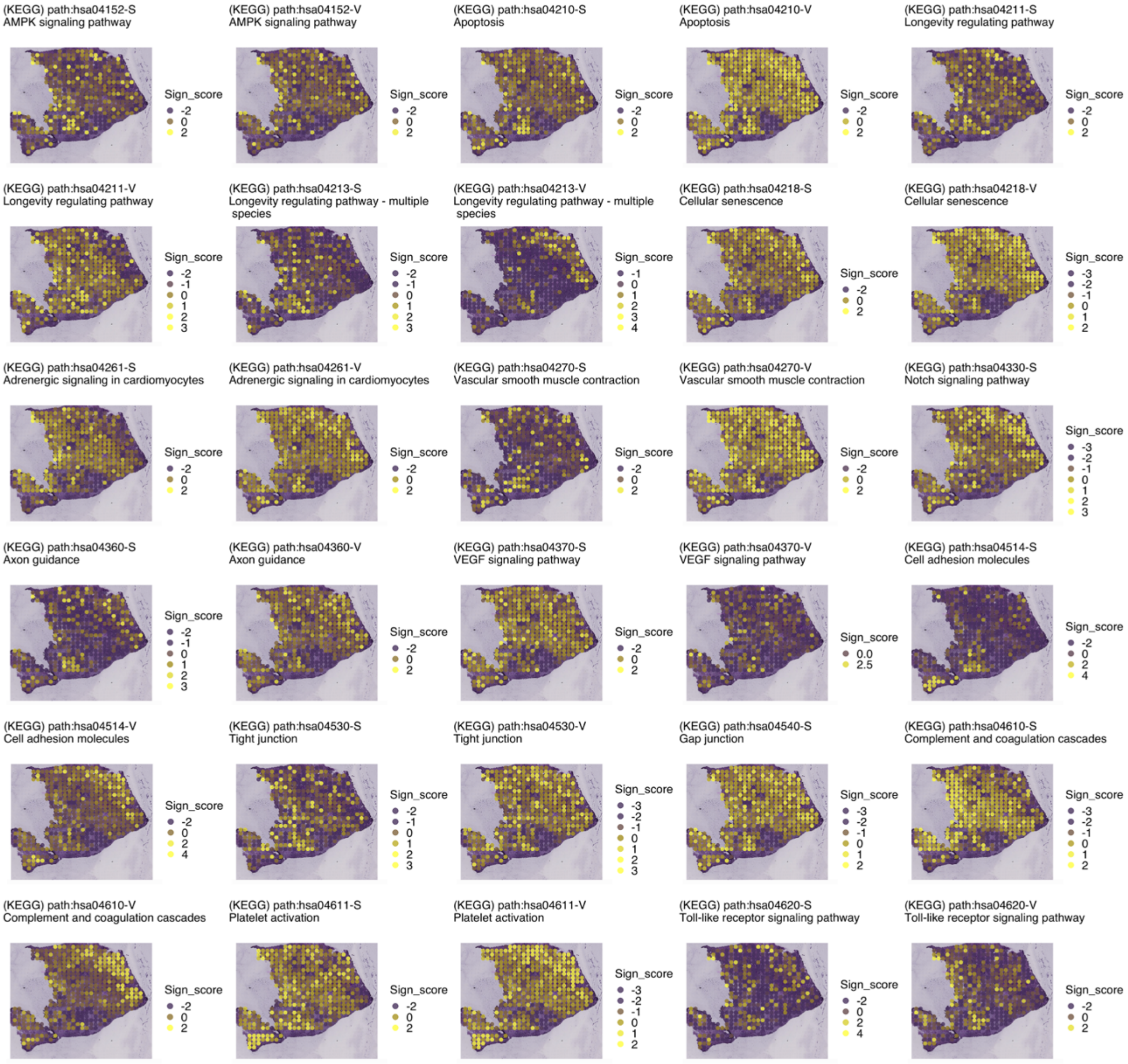
(GO:BP) GO:0071347-S
Cellular response to interleukin-1


(GO:BP) GO:1902476-S
Chloride transmembrane transport
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0033108-S
Mitochondrial respiratory chain complex assembly
Sign_score 0.0 2.5 5.0
(GO:BP) GO:0043624-S
Cellular protein complex disassembly
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0007044-S
Cell-substrate junction assembly
Sign_score -2 -1 0 1 2
(GO:BP) GO:0031529-S
Ruffle organization
Sign_score -3 -2 -1 0 1 2
(GO:BP) GO:0006376-S
mRNA splice site selection
Sign_score -2.5 0.0 2.5 5.0
(GO:BP) GO:0060048-S
Cardiac muscle contraction
Sign_score -2 0 2
(GO:BP) GO:0006953-S
Acute-phase response
Sign_score -2 -1 0 1 2
(GO:BP) GO:0051937-S
Catecholamine transport
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0006367-S
Transcription initiation from RNA polymerase II promoter
Sign_score -2 0 2
(GO:BP) GO:0032515-S
Negative regulation of phosphoprotein phosphatase activity
Sign_score -2 0 2
(GO:BP) GO:0031397-S
Negative regulation of protein ubiquitination
Sign_score -2 0 2 4
(GO:BP) GO:0031397-V
Negative regulation of protein ubiquitination
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:1902895-S
Positive regulation of pri-miRNA transcription by RNA polymerase II
Sign_score -2 0 2
(GO:BP) GO:0018105-S
Peptidyl-serine phosphorylation
Sign_score -2.5 0.0 2.5
(GO:BP) GO:0018105-V
Peptidyl-serine phosphorylation
Sign_score -2 0 2 4
(GO:BP) GO:0006303-S
Double-strand break repair via nonhomologous end joining
Sign_score -2 0 2 4
(GO:BP) GO:0006283-S
Transcription-coupled nucleotide-excision repair
Sign_score -2 0 2 4
(GO:BP) GO:0006283-V
Transcription-coupled nucleotide-excision repair
Sign_score -4 -2 0 2
(GO:BP) GO:0032435-S
Negative regulation of proteasomal ubiquitin-dependent protein catabolic...
Sign_score -2 0 2
(GO:BP) GO:0031124-S
mRNA 3'-end processing
Sign_score -2 0 2 4
(GO:BP) GO:0007173-S
Epidermal growth factor receptor signaling pathway
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0046329-S
Negative regulation of JNK cascade
Sign_score -2 0 2
(GO:BP) GO:2000352-S
Negative regulation of endothelial cell apoptotic process
Sign_score -2 0 2 4
(GO:BP) GO:0006882-S
Cellular zinc ion homeostasis
Sign_score -2 0 2 4
(GO:BP) GO:0046323-S
Glucose import
Sign_score -1 0 1 2 3 4
(GO:BP) GO:0046323-V
Glucose import
Sign_score -2 0 2
(GO:BP) GO:0030032-S
Lamellipodium assembly
Sign_score -2 -1 0 1 2 3
(GO:BP) GO:0009749-S
Response to glucose
Sign_score -2 0 2 4
(GO:BP) GO:0009749-V
Response to glucose
Sign_score -2 -1 0 1 2 3


(GO:BP) GO:0006813-S
Potassium ion transport
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0061098-S
Positive regulation of protein tyrosine kinase activity
Sign_score
-2
-1
0
1
2
(GO:BP) GO:0000413-S
Protein peptidyl-prolyl isomerization
Sign_score
-2
-1
0
1
2
3
(GO:BP) GO:0000377-S
RNA splicing, via transesterification reactions with bulged adenosine as...
Sign_score
-2.5
0.0
2.5
(GO:BP) GO:0000377-V
RNA splicing, via transesterification reactions with bulged adenosine as...
Sign_score
-2
0
2
(GO:BP) GO:0055078-S
Sodium ion homeostasis
Sign_score
-2
0
2


(KEGG) path:hsa00051-S Fructose and mannose metabolism
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00052-S Galactose metabolism
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00230-S Purine metabolism
Sign_score -2 0 2 4
(KEGG) path:hsa00240-S Pyrimidine metabolism
Sign_score -2 0 2
(KEGG) path:hsa00240-V Pyrimidine metabolism
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00480-S Glutathione metabolism
Sign_score -2 0 2 4
(KEGG) path:hsa00480-V Glutathione metabolism
Sign_score -2.5 0.0 2.5
(KEGG) path:hsa00510-S N-Glycan biosynthesis
Sign_score -2 0 2
(KEGG) path:hsa00510-V N-Glycan biosynthesis
Sign_score -2.5 0.0 2.5
(KEGG) path:hsa00590-S Arachidonic acid metabolism
Sign_score -2 0 2 4
(KEGG) path:hsa00600-S Sphingolipid metabolism
Sign_score -2 0 2
(KEGG) path:hsa00630-S Glyoxylate and dicarboxylate metabolism
Sign_score -2 0 2
(KEGG) path:hsa00980-S Metabolism of xenobiotics by cytochrome P450
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00980-V Metabolism of xenobiotics by cytochrome P450
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00982-S Drug metabolism - cytochrome P450
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00982-V Drug metabolism - cytochrome P450
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00983-S Drug metabolism - other enzymes
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa00983-V Drug metabolism - other enzymes
Sign_score -2 0 2 4
(KEGG) path:hsa01524-S Platinum drug resistance
Sign_score -2 0 2
(KEGG) path:hsa01524-V Platinum drug resistance
Sign_score -2 0 2
(KEGG) path:hsa03008-S Ribosome biogenesis in eukaryotes
Sign_score -2 0 2
(KEGG) path:hsa03010-S Ribosome
Sign_score -2 0 2 4
(KEGG) path:hsa03010-V Ribosome
Sign_score -2.5 0.0 2.5
(KEGG) path:hsa03018-S RNA degradation
Sign_score -2 0 2
(KEGG) path:hsa03020-S RNA polymerase
Sign_score -2 0 2
(KEGG) path:hsa03040-S Spliceosome
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa03040-V Spliceosome
Sign_score -2 0 2
(KEGG) path:hsa04024-S cAMP signaling pathway
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa04024-V cAMP signaling pathway
Sign_score -3 -2 -1 0 1 2 3
(KEGG) path:hsa04060-S Cytokine-cytokine receptor interaction
Sign_score -2 0 2 4


(KEGG) path:hsa04060-V
Cytokine-cytokine receptor interaction
Sign_score
-2
0
2
4
(KEGG) path:hsa04062-S
Chemokine signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04062-V
Chemokine signaling pathway
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04064-S
NF-kappa B signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04064-V
NF-kappa B signaling pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04068-S
FoxO signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04068-V
FoxO signaling pathway
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04070-S
Phosphatidylinositol signaling system
Sign_score
-2
0
2
(KEGG) path:hsa04070-V
Phosphatidylinositol signaling system
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04071-S
Sphingolipid signaling pathway
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04071-V
Sphingolipid signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04072-S
Phospholipase D signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04072-V
Phospholipase D signaling pathway
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04110-S
Cell cycle
Sign_score
-2
0
2
4
(KEGG) path:hsa04110-V
Cell cycle
Sign_score
-2
0
2
4
(KEGG) path:hsa04115-S
p53 signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04115-V
p53 signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04120-S
Ubiquitin mediated proteolysis
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04120-V
Ubiquitin mediated proteolysis
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04130-S
SNARE interactions in vesicular transport
Sign_score
-2
0
2
4
(KEGG) path:hsa04140-S
Autophagy - animal
Sign_score
-2
0
2
4
(KEGG) path:hsa04140-V
Autophagy - animal
Sign_score
-2
0
2
(KEGG) path:hsa04144-S
Endocytosis
Sign_score
-2
0
2
(KEGG) path:hsa04144-V
Endocytosis
Sign_score
-2
0
2
(KEGG) path:hsa04145-S
Phagosome
Sign_score
-2
0
2
4
(KEGG) path:hsa04145-V
Phagosome
Sign_score
-2
0
2
(KEGG) path:hsa04150-S
mTOR signaling pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04150-V
mTOR signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04151-S
PI3K-Akt signaling pathway
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04151-V
PI3K-Akt signaling pathway
Sign_score
-2
-1
0
1
2
3



(KEGG) path:hsa04152-S
AMPK signaling pathway

(KEGG) path:hsa04152-V
AMPK signaling pathway

(KEGG) path:hsa04210-S
Apoptosis

(KEGG) path:hsa04210-V
Apoptosis

(KEGG) path:hsa04211-S
Longevity regulating pathway

(KEGG) path:hsa04211-V
Longevity regulating pathway

(KEGG) path:hsa04213-S
Longevity regulating pathway - multiple species

(KEGG) path:hsa04213-V
Longevity regulating pathway - multiple species

(KEGG) path:hsa04218-S
Cellular senescence

(KEGG) path:hsa04218-V
Cellular senescence

(KEGG) path:hsa04261-S
Adrenergic signaling in cardiomyocytes

(KEGG) path:hsa04261-V
Adrenergic signaling in cardiomyocytes

(KEGG) path:hsa04270-S
Vascular smooth muscle contraction

(KEGG) path:hsa04270-V
Vascular smooth muscle contraction

(KEGG) path:hsa04330-S
Notch signaling pathway

(KEGG) path:hsa04360-S
Axon guidance

(KEGG) path:hsa04360-V
Axon guidance

(KEGG) path:hsa04370-S
VEGF signaling pathway

(KEGG) path:hsa04370-V
VEGF signaling pathway

(KEGG) path:hsa04514-S
Cell adhesion molecules

(KEGG) path:hsa04514-V
Cell adhesion molecules

(KEGG) path:hsa04530-S
Tight junction

(KEGG) path:hsa04530-V
Tight junction

(KEGG) path:hsa04540-S
Gap junction

(KEGG) path:hsa04610-S
Complement and coagulation cascades

(KEGG) path:hsa04610-V
Complement and coagulation cascades

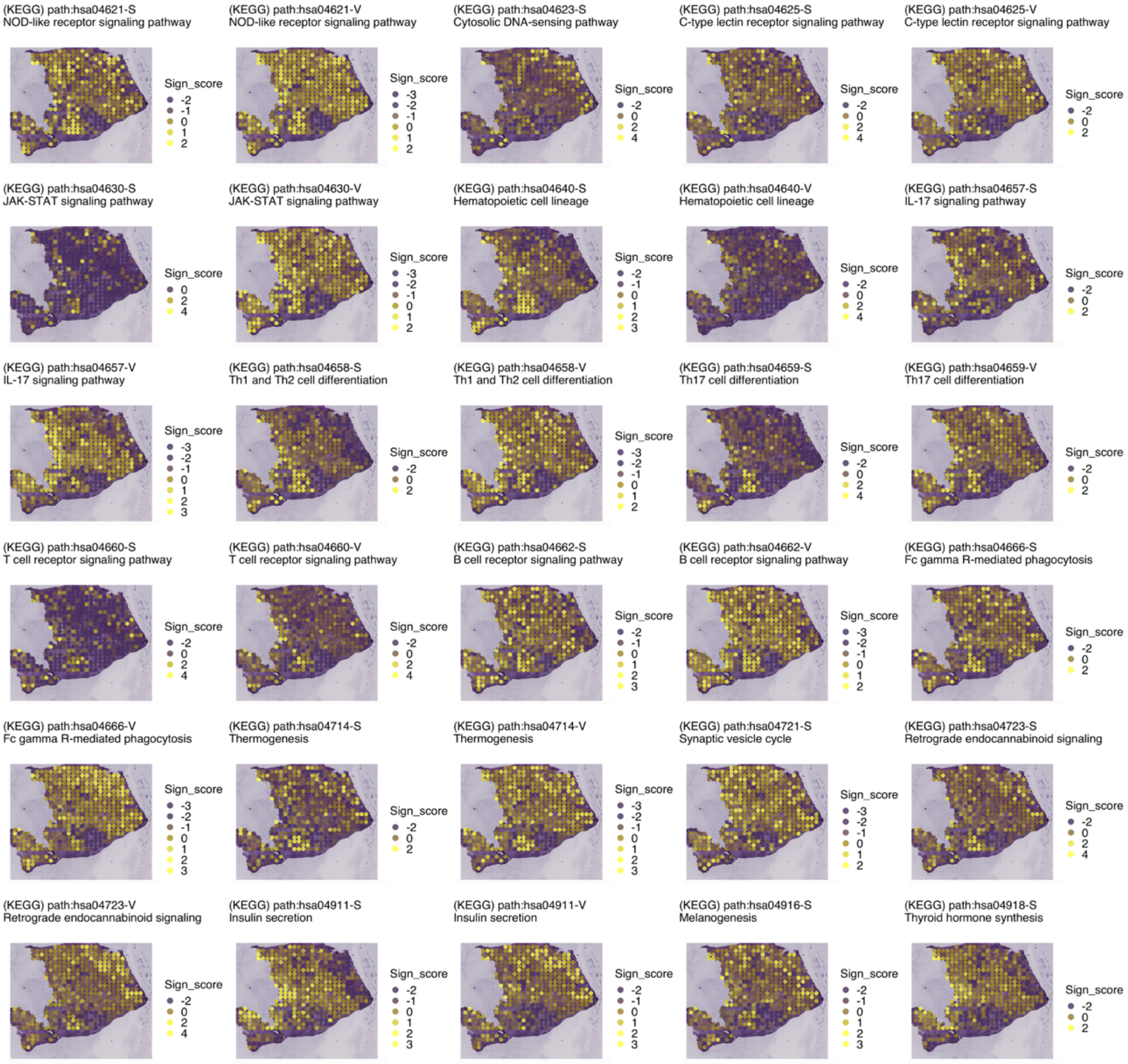
(KEGG) path:hsa04611-S
Platelet activation

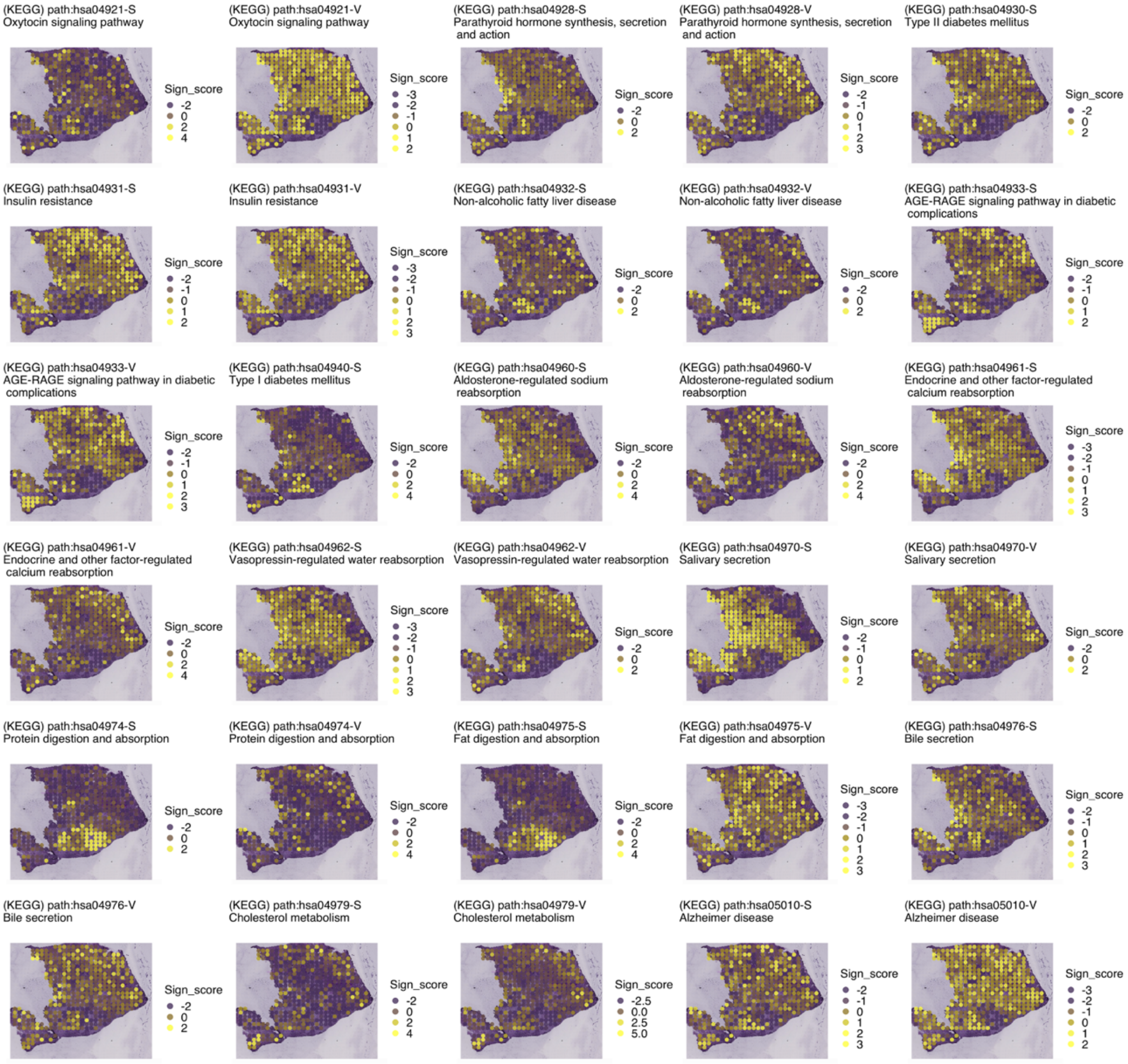
(KEGG) path:hsa04611-V
Platelet activation

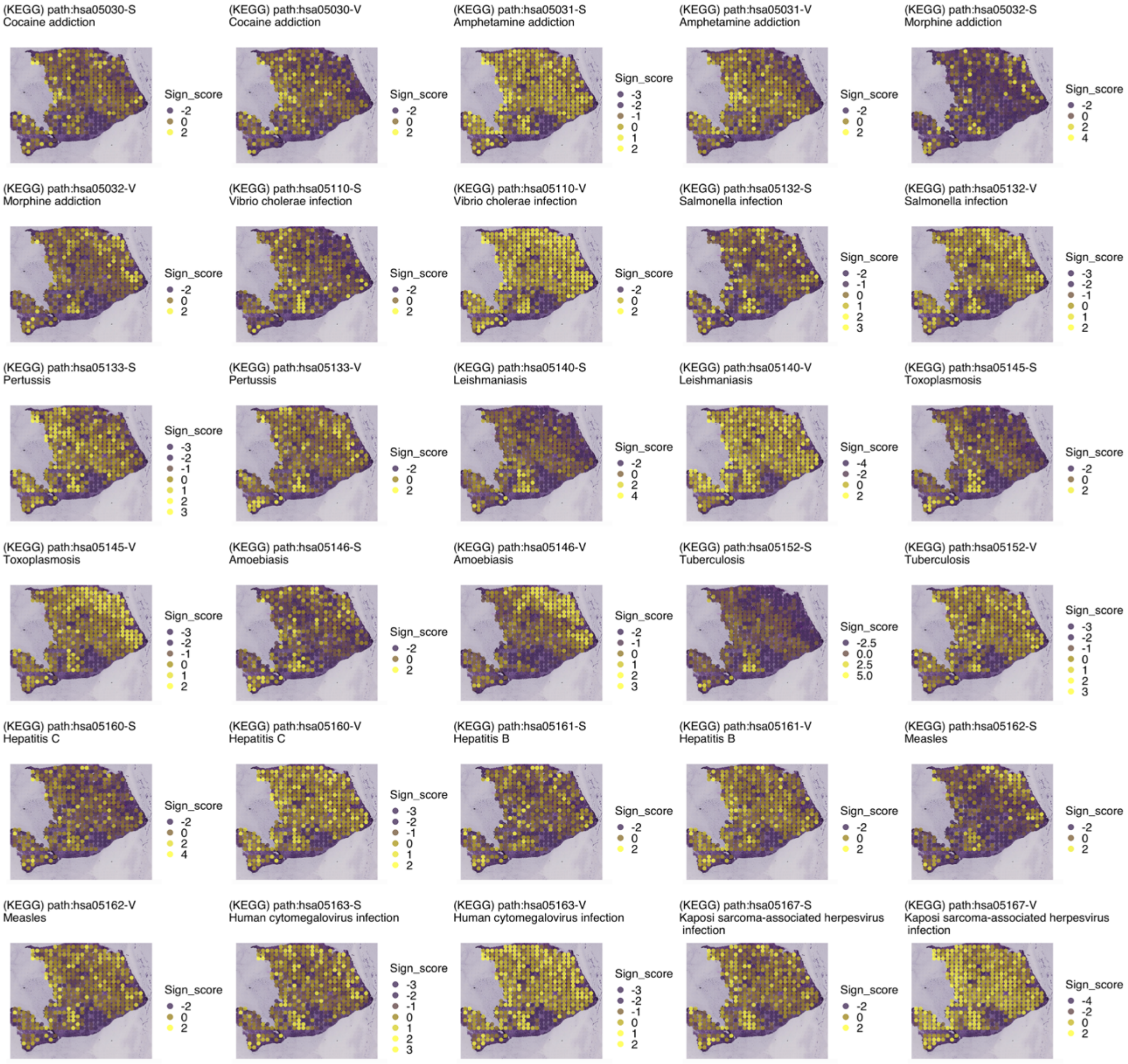
(KEGG) path:hsa04620-S
Toll-like receptor signaling pathway

(KEGG) path:hsa04620-V
Toll-like receptor signaling pathway

Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● -1 ● 0 ● 1 ● 2 ● 3
Sign_score ● -2 ● -1 ● 0 ● 1 ● 2 ● 3
Sign_score ● -1 ● 0 ● 1 ● 2 ● 3 ● 4
Sign_score ● -2 ● 0 ● 2
Sign_score ● -3 ● -2 ● -1 ● 0 ● 1 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -3 ● -2 ● -1 ● 0 ● 1 ● 2 ● 3
Sign_score ● -2 ● -1 ● 0 ● 1 ● 2 ● 3
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● 0.0 ● 2.5
Sign_score ● -2 ● 0 ● 2 ● 4
Sign_score ● -2 ● 0 ● 2 ● 4
Sign_score ● -2 ● -1 ● 0 ● 1 ● 2 ● 3
Sign_score ● -3 ● -2 ● -1 ● 0 ● 1 ● 2 ● 3
Sign_score ● -3 ● -2 ● -1 ● 0 ● 1 ● 2
Sign_score ● -3 ● -2 ● -1 ● 0 ● 1 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -2 ● 0 ● 2
Sign_score ● -3 ● -2 ● -1 ● 0 ● 1 ● 2
Sign_score ● -2 ● 0 ● 2 ● 4
Sign_score ● -2 ● 0 ● 2


(KEGG) path:hsa04621-S
NOD-like receptor signaling pathway
Sign_score
-2
-1
0
1
2
(KEGG) path:hsa04621-V
NOD-like receptor signaling pathway
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04623-S
Cytosolic DNA-sensing pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04625-S
C-type lectin receptor signaling pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04625-V
C-type lectin receptor signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04630-S
JAK-STAT signaling pathway
Sign_score
0
2
4
(KEGG) path:hsa04630-V
JAK-STAT signaling pathway
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04640-S
Hematopoietic cell lineage
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04640-V
Hematopoietic cell lineage
Sign_score
-2
0
2
4
(KEGG) path:hsa04657-S
IL-17 signaling pathway
Sign_score
-2
0
2
(KEGG) path:hsa04657-V
IL-17 signaling pathway
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04658-S
Th1 and Th2 cell differentiation
Sign_score
-2
0
2
(KEGG) path:hsa04658-V
Th1 and Th2 cell differentiation
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04659-S
Th17 cell differentiation
Sign_score
-2
0
2
4
(KEGG) path:hsa04659-V
Th17 cell differentiation
Sign_score
-2
0
2
(KEGG) path:hsa04660-S
T cell receptor signaling pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04660-V
T cell receptor signaling pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04662-S
B cell receptor signaling pathway
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04662-V
B cell receptor signaling pathway
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04666-S
Fc gamma R-mediated phagocytosis
Sign_score
-2
0
2
(KEGG) path:hsa04666-V
Fc gamma R-mediated phagocytosis
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04714-S
Thermogenesis
Sign_score
-2
0
2
(KEGG) path:hsa04714-V
Thermogenesis
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04721-S
Synaptic vesicle cycle
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04723-S
Retrograde endocannabinoid signaling
Sign_score
-2
0
2
4
(KEGG) path:hsa04723-V
Retrograde endocannabinoid signaling
Sign_score
-2
0
2
4
(KEGG) path:hsa04911-S
Insulin secretion
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04911-V
Insulin secretion
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04916-S
Melanogenesis
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04918-S
Thyroid hormone synthesis
Sign_score
-2
0
2


(KEGG) path:hsa04921-S
Oxytocin signaling pathway
Sign_score
-2
0
2
4
(KEGG) path:hsa04921-V
Oxytocin signaling pathway
Sign_score
-3
-2
-1
0
1
2
(KEGG) path:hsa04928-S
Parathyroid hormone synthesis, secretion and action
Sign_score
-2
0
2
(KEGG) path:hsa04928-V
Parathyroid hormone synthesis, secretion and action
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04930-S
Type II diabetes mellitus
Sign_score
-2
0
2
(KEGG) path:hsa04931-S
Insulin resistance
Sign_score
-2
-1
0
1
2
(KEGG) path:hsa04931-V
Insulin resistance
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04932-S
Non-alcoholic fatty liver disease
Sign_score
-2
0
2
(KEGG) path:hsa04932-V
Non-alcoholic fatty liver disease
Sign_score
-2
0
2
(KEGG) path:hsa04933-S
AGE-RAGE signaling pathway in diabetic complications
Sign_score
-2
-1
0
1
2
(KEGG) path:hsa04933-V
AGE-RAGE signaling pathway in diabetic complications
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04940-S
Type I diabetes mellitus
Sign_score
-2
0
2
4
(KEGG) path:hsa04960-S
Aldosterone-regulated sodium reabsorption
Sign_score
-2
0
2
4
(KEGG) path:hsa04960-V
Aldosterone-regulated sodium reabsorption
Sign_score
-2
0
2
4
(KEGG) path:hsa04961-S
Endocrine and other factor-regulated calcium reabsorption
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04961-V
Endocrine and other factor-regulated calcium reabsorption
Sign_score
-2
0
2
4
(KEGG) path:hsa04962-S
Vasopressin-regulated water reabsorption
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04962-V
Vasopressin-regulated water reabsorption
Sign_score
-2
0
2
(KEGG) path:hsa04970-S
Salivary secretion
Sign_score
-2
-1
0
1
2
(KEGG) path:hsa04970-V
Salivary secretion
Sign_score
-2
0
2
(KEGG) path:hsa04974-S
Protein digestion and absorption
Sign_score
-2
0
2
(KEGG) path:hsa04974-V
Protein digestion and absorption
Sign_score
-2
0
2
4
(KEGG) path:hsa04975-S
Fat digestion and absorption
Sign_score
-2
0
2
4
(KEGG) path:hsa04975-V
Fat digestion and absorption
Sign_score
-3
-2
-1
0
1
2
3
(KEGG) path:hsa04976-S
Bile secretion
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa04976-V
Bile secretion
Sign_score
-2
0
2
(KEGG) path:hsa04979-S
Cholesterol metabolism
Sign_score
-2
0
2
4
(KEGG) path:hsa04979-V
Cholesterol metabolism
Sign_score
-2.5
0.0
2.5
5.0
(KEGG) path:hsa05010-S
Alzheimer disease
Sign_score
-2
-1
0
1
2
3
(KEGG) path:hsa05010-V
Alzheimer disease
Sign_score
-3
-2
-1
0
1
2


(KEGG) path:hsa05030-S
Cocaine addiction
Sign_score -2 0 2
(KEGG) path:hsa05030-V
Cocaine addiction
Sign_score -2 0 2
(KEGG) path:hsa05031-S
Amphetamine addiction
Sign_score -3 -2 -1 0 1 2
(KEGG) path:hsa05031-V
Amphetamine addiction
Sign_score -2 0 2
(KEGG) path:hsa05032-S
Morphine addiction
Sign_score -2 0 2 4
(KEGG) path:hsa05032-V
Morphine addiction
Sign_score -2 0 2
(KEGG) path:hsa05110-S
Vibrio cholerae infection
Sign_score -2 0 2
(KEGG) path:hsa05110-V
Vibrio cholerae infection
Sign_score -2 0 2
(KEGG) path:hsa05132-S
Salmonella infection
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa05132-V
Salmonella infection
Sign_score -3 -2 -1 0 1 2
(KEGG) path:hsa05133-S
Pertussis
Sign_score -3 -2 -1 0 1 2 3
(KEGG) path:hsa05133-V
Pertussis
Sign_score -2 0 2
(KEGG) path:hsa05140-S
Leishmaniasis
Sign_score -2 0 2 4
(KEGG) path:hsa05140-V
Leishmaniasis
Sign_score -4 -2 0 2
(KEGG) path:hsa05145-S
Toxoplasmosis
Sign_score -2 0 2
(KEGG) path:hsa05145-V
Toxoplasmosis
Sign_score -3 -2 -1 0 1 2
(KEGG) path:hsa05146-S
Amoebiasis
Sign_score -2 0 2
(KEGG) path:hsa05146-V
Amoebiasis
Sign_score -2 -1 0 1 2 3
(KEGG) path:hsa05152-S
Tuberculosis
Sign_score -2.5 0.0 2.5 5.0
(KEGG) path:hsa05152-V
Tuberculosis
Sign_score -3 -2 -1 0 1 2 3
(KEGG) path:hsa05160-S
Hepatitis C
Sign_score -2 0 2 4
(KEGG) path:hsa05160-V
Hepatitis C
Sign_score -3 -2 -1 0 1 2
(KEGG) path:hsa05161-S
Hepatitis B
Sign_score -2 0 2
(KEGG) path:hsa05161-V
Hepatitis B
Sign_score -2 0 2
(KEGG) path:hsa05162-S
Measles
Sign_score -2 0 2
(KEGG) path:hsa05162-V
Measles
Sign_score -2 0 2
(KEGG) path:hsa05163-S
Human cytomegalovirus infection
Sign_score -3 -2 -1 0 1 2 3
(KEGG) path:hsa05163-V
Human cytomegalovirus infection
Sign_score -3 -2 -1 0 1 2
(KEGG) path:hsa05167-S
Kaposi sarcoma-associated herpesvirus infection
Sign_score -2 0 2
(KEGG) path:hsa05167-V
Kaposi sarcoma-associated herpesvirus infection
Sign_score -4 -2 0 2

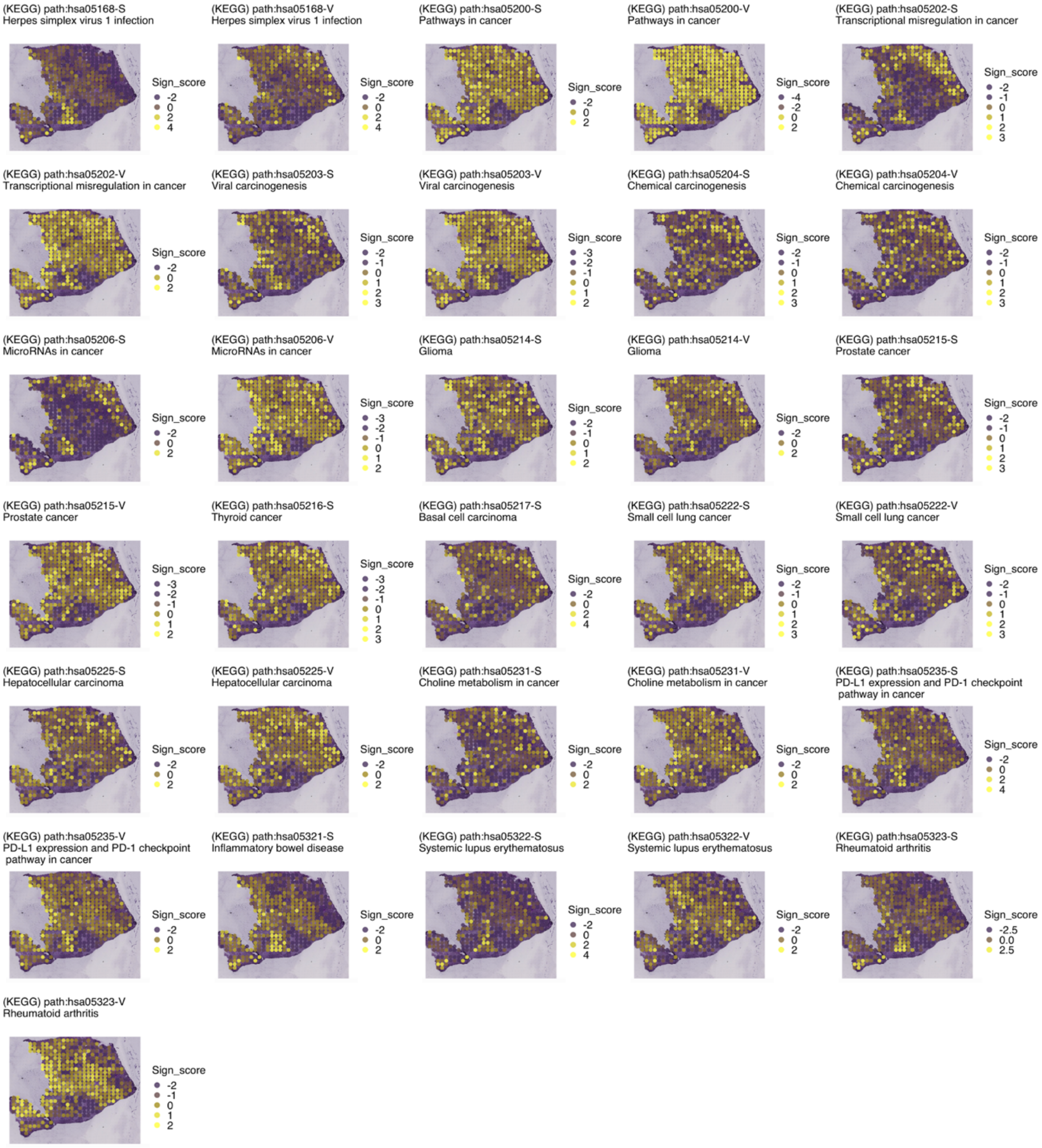


Figure S6. Sign scores for functions and signaling pathway activities using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) across the pancreatic ductal adenocarcinoma (PDAC) tissue.

Supplementary Note 1. Parameter settings of ASURAT

To obtain desired results, it is critical to tune ASURAT parameters for creating sign-by-sample matrices (SSMs). Depending on the DBs, there were six to nine parameters for creating SSMs, but many of them have been preset to unbiased and sensible default values (Figure S1). We found that our default settings worked well in our single-cell RNA sequencing (scRNA-seq) analyses, but the three parameters should be tuned by users, as described below.

As formulated in **Error! Reference source not found.**, positive and negative constants α and β from thresholds of correlation coefficients are required for decomposing correlation graphs and creating signs (see **Figure 2** for the demonstration). In addition, unreliable signs are discarded with user-defined criteria, which were preset as follows: the sum of the number of genes in the strongly and variably correlated gene sets, SCG and VCG, respectively, is less than n_{min} or the number of genes in weakly correlated gene set (WCG) is less than $n_{\mathrm{min}}^{(\mathrm{w})}$ (the default value is 2). Furthermore, users can remove redundant signs with similar biological meanings if information contents (ICs) (Yu, et al., 2010) are defined.

Supplementary Note 2. Separation index

Briefly, a separation index is a measure of significance of a given sign score for a given subpopulation. Since the row vectors of SSMs are centered (i.e., the means are zeros), wherein the degree of freedom is reduced, naïve usages of statistical tests and fold change analyses should be avoided. Nevertheless, we propose helping users to find significant signs using a nonparametric index to quantify the extent of separation between two sets of random variables. A separation index of a given random variable *X* takes a value from -1 to 1: the larger positive value indicates that *X*s are markedly upregulated, and the probability distribution is well separated against other distributions and vice versa.

Let us consider a vector $\boldsymbol{a}$ of size n, i.e., the number of samples, whose elements stand for the sign scores, and assume that the elements are sorted in ascending order. For simplicity suppose that the samples are classified into two clusters labeled 0 and 1. Let $\boldsymbol{v}$ be a vector of the labels corresponding to $\boldsymbol{a}$, and $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ be vectors having the same elements with $\boldsymbol{v}$ but the elements are sorted in lexicographic orders in forward and backward directions, respectively. Then, we define the separation index as follows:

$$I(\boldsymbol{v}) = 1 - \frac{2d(\boldsymbol{v},\boldsymbol{w}_0)}{d(\boldsymbol{v},\boldsymbol{w}_0)+d(\boldsymbol{v},\boldsymbol{w}_1)}, \tag{3}$$

where $d(\boldsymbol{v}, \boldsymbol{w}_i)$ is an edit distance (or Levenshtein distance (Lowrance and Wagner, 1975)) with only adjacent swapping permitted. For example, if $\boldsymbol{v} = (1, 0, 0, 1, 1)$, then $\boldsymbol{w}_0 = (0, 0, 1, 1, 1)$ and $\boldsymbol{w}_1 = (1, 1, 1, 0, 0)$. From (3) one can calculate $d(\boldsymbol{v}, \boldsymbol{w}_0) = 2$ and $d(\boldsymbol{v}, \boldsymbol{w}_1) = 4$, and thus $I(\boldsymbol{v}) = 1/3$. As another example, if $\boldsymbol{v} = (0, 1, 1, 0, 0)$, then $I(\boldsymbol{v}) = -1/3$. From this example, one can see that the positive and negative values of I mean that the given sign has positive and negative contributions for cluster "1," respectively.

Supplementary Note 3. Datasets

Human peripheral blood mononuclear cells

These data were obtained from peripheral blood mononuclear cells (PBMCs) of healthy donors, which include approximately 4,000 and 6,000 cells; thus, they were referred to as PBMCs 4k and 6k, respectively. The data were produced with a 10x protocol using unique molecular identifiers (UMIs). The single-cell transcriptome datasets were downloaded from the 10x Genomics repository (https://support.10xgenomics.com/single-cell-gene-expression/datasets). The following filtered read count matrices were obtained: PBMC 4k from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k) and PBMC 6k from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k). After data quality controls, the read count tables of PBMC 4k (resp. PBMC 6k) contained 6,658 (resp. 5,169) genes and 3,815 (resp. 4,878) cells.

Human small cell lung cancer with cisplatin treatments

The data were obtained from circulating tumor cell-derived xenografts cultured with cisplatin treatments, which were generated from lung cancer patients (Stewart, et al., 2020). The data were produced with a 10x protocol using UMIs. The SRA files were downloaded from Gene Expression Omnibus (GEO) with accession codes GSE138474: GSM4104164, which is referenced in Stewart et al. (2020). SRA Toolkit version 2.10.8 was used to dump the FASTQ files. Cell Ranger version 3.1.0 was used to align the FASTQ files to the GRCh38-3.0.0 human reference genome and produce the single-cell transcriptome datasets. After controlling for data quality, the read count table contained 6,581 genes and 3,923 cells.

Human pancreatic ductal adenocarcinoma

The single-cell RNA sequencing (scRNA-seq) and spatial transcriptome (ST) data were obtained from PDAC patients using inDrop and ST protocols (Moncada, et al., 2020), respectively. The FASTQ files were downloaded from Gene Expression Omnibus (GEO) with accession codes GSE111672: GSM3036909, GSM3036910, GSM3036911, GSM3405527, GSM3405528, GSM3405529, and GSM3405530. Mapping of raw sequencing data from inDrop and ST protocols were processed using custom pipelines from https://github.com/flo-compbio/singlecell and https://github.com/jfnavarro/st_pipeline, respectively. Both pipelines used the parameters explained by Moncada et al. (2020). Prior to downstream analysis, we concatenated all the scRNA-seq datasets. After data quality controls, the read count table of the combined scRNA-seq dataset contained 5,893 genes and 2,051 cells, wherein the ST dataset contained 4,497 genes and 428 ST spots. ST data was imported and visualized using Spaniel (Queen, et al., 2019).

Supplementary Note 4. Data preprocessing: quality control, normalization, and centering

For all the scRNA-seq datasets, the low-quality genes and cells were removed by the following three steps: (i) removing the genes for which the number of non-zero expressing cells is less than a user-defined threshold; (ii) removing the cells whose read counts, number of genes expressed with non-zero read counts, and percent of reads mapped to mitochondrial genes are within user-defined ranges; and (iii) removing the genes for which the mean of the read counts is less than a user-defined threshold. See Chapters 2 and 3 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

After applying data quality controls, the data were normalized by bayNorm (Tang, et al., 2020), which attenuates technical biases with respect to zero inflation and variation of capture efficiencies between cells. The resulting inferred true count matrices were supplied to a log-transformation with a pseudo-count to attenuate the impact of dispersion in the counts for highly expressed genes. Finally, subtracting the sample mean from each row vector, we obtained the normalized-and-centered read count tables. See Chapter 4 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Supplementary Note 5. Analysis of scRNA-seq datasets of PBMC 4k and 6k

To compare the cell-type inference abilities of existing methods and ASURAT, we prepared two scRNA-seq datasets, namely PBMCs 4k and 6k (see Datasets). Subsequently, data quality controls and normalization by bayNorm were carefully performed for each dataset. See Chapters 2–4 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Using scran (version 1.18.7) (Lun, et al., 2016), we normalized the data using the functions quickCluster(), computeSumFactors(), and logNormCounts(), selected highly variable genes using modelGeneVar() and getTopHVGs() based on a variance modeling with a gene-per-cell ratio of 0.2 (as suggested in a previous work (Cruz and Wishart, 2007)), and set the principal components using denoisePCA(). Cells were clustered using buildSNNGraph() and cluster_louvain(). Then, candidates of differentially expressed genes (DEGs) were detected using pairwiseTTests() and combineMarkers(), and DEGs were defined as genes with false discovery rates (FDRs)$< 10^{-99}$ (T tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer, et al., 2016) as follows: B cells (resp. marker genes *CD79A*, *MS4A1*, *IGHM*), monocytes (*S100A8*, *LYZ*, *CD14*), NK or NKT cells (*NKG7*, *GZMA*, *FGFBP2*), and T cells (*MAL*). See Chapter 13 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Using Seurat (version 4.0.2) (Hao, et al., 2021), we normalized the data using the function NormalizeData() with a log normalization (default), selected highly variable genes using

FindVariableFeatures() based on a variance-stabilizing transformation with a gene-per-cell ratio of 0.2 (as suggested in previous work (Cruz and Wishart, 2007)), scaled and centered gene expression levels, and performed PCA. The principal components that explained 90% of the total variability were used for the computations of FindNeighbors(). Cells were clustered using FindClusters(). Then, candidates of DEGs were detected using FindAllMarkers() and DEGs were defined as genes with false discovery rates (FDRs)$<$ 10^{-99} (Mann-Whitney *U* tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer, et al., 2016) as follows: T cells (resp. marker genes *TRAC*, *CD3D*, *IL32*, *TCF7*, *CD27*), monocytes (*S100A8*, *LYZ*, *CD14*), B cells (*CD79A*, *MS4A1*, *IGHM*, *VPREB3*, *BANK1*), and NK or NKT cells (*CD3D*, *NKG7*, *GZMA*, *FGFBP2*). Additionally, to automatically annotate the clustering results, we used the R function findmarkergenes() in the scCATCH (version 2.1) package (Shao, et al., 2020), which identified monocytes, B cells, and T cells. See Chapter 14 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Using Monocle 3 (version 1.0.0) (Trapnell, et al., 2014), we used the function preprocess_cds() under the default settings, in which data were normalized by a log transform with a pseudo-count of 1, scaled and centered in gene expression levels, and were subjected to PCA with the dimensionality of the reduced space set to 50. Cells were clustered by cluster_cells() using Uniform Manifold Approximation and Projection (UMAP) (McInnes and Healy, 2018). Then, candidate DEGs were detected using top_markers() and DEGs were defined as genes with false discovery rates (FDRs)$<$ 10^{-99} (Monocle's marker significance tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer, et al., 2016) as follows: T cells (resp. marker genes *CD3D*, *TCF7*, *CD3E*, *IL32*), monocytes (*S100A8*, *LYZ*, *CD14*), B cells (*CD79A*, *CD79B*, *BANK1*, *MS4A1*), and NK or NKT cells (*GNLY*, *NKG7*, *GZMA*). See Chapter 15 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Using SC3 (version 1.18.0) (Kiselev, et al., 2017), we performed the function runPCA() inputting log-normalized read count tables with a pseudo-count of 1. Cells were clustered using sc3(), and reasonable numbers of clusters were manually determined by sc3_plot_markers(). Then, candidate DEGs were detected using get_marker_genes() and DEGs were defined as genes with false discovery rates (FDRs)$< 10^{-99}$ (Kruskal-Wallis tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer, et al., 2016) as follows: NK or NKT cells (resp. marker genes *GZMA*, *GZMB*, *GZMH*, *GZMK*, *GNLY*), T cells (*TRGC2*, *TCL1A*), monocytes (*GSN*, *LILRB4*, *S100A8*, *CD14*, *S100A12*), and B cells (*CD79A*, *CD79B*, *MS4A1*, *SPI1*, *LYN*). See Chapter 16 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Using ASURAT, we created SSMs using the CO, GO, and KEGG DBs. After dimensionality reduction by PCA, cells were clustered by *k*-nearest neighbor (KNN) graph generation and Louvain algorithm using Seurat functions FindNeighbors() and FindClusters() (Hao, et al., 2021). Subsequently, separation indices (SIs) were computed for all the signs for a given cluster versus all the others, then cell types were identified by manually selecting significant signs with the larger values of SIs$>$ 0.5 (**Figure 3**). See Chapter 17 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Supplementary Note 6. Analysis of an SCLC scRNA-seq dataset

For the analysis of an SCLC scRNA-seq dataset, we began the Seurat workflow by normalizing data using the Seurat function NormalizeData() with a log normalization (default). Then, highly variable genes were selected by FindVariableFeatures() based on a variance stabilizing transformation with a gene-per-cell ratio of 0.2 (as suggested in previous work (Cruz and Wishart, 2007)). Then, data were scaled and centered by ScaleData(), and PCA was applied by RunPCA() with highly variable genes. Subsequently, a KNN graph was generated by FindNeighbors(), with the principal components that explain 90% of the total variability, and cells were clustered by FindClusters() with a Louvain algorithm. Additionally, cell cycle phases were inferred by CellCycleScoring() with cell cycle-related genes defined in the Seurat package. Finally, KEGG enrichment analysis was done by compareCluster() in clusterProfiler package (Yu, et al., 2012). See Chapter 14 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

The ASURAT workflow started with the collection of DO, GO, and KEGG databases. First, we excluded functional gene sets including too few or too many genes. Next, we created multiple signs using a correlation graph-based decomposition. Then, we removed redundant signs with similar biological meanings using doSim() from the DOSE package (Yu, et al., 2015). We then created SSMs for DO, GO, and KEGG. Based on the SSM for DO, we performed a dimensionality reduction using the diffusion map and clustered cells using MERLoT (Parra, et al., 2019). Finally, we vertically concatenated all the SSMs, cell cycle phases inferred by Seurat, and expression matrix for characterizing individual cells from multiple biological aspects. The DEGs were identified using FindAllMarkers() in Seurat package. See Chapters 9–12 in our tutorial (https://keita-iida.github.io/ASURAT_0.0.0.9001/index.html).

Supplementary Note 7. Limitations of the study

To formulate signs, we used a correlation graph-based decomposition based on functional gene sets (FGSs) with thresholds set as positive and negative correlation coefficients (**Figure 2**), from which we obtained SCGs, VCGs, and WCGs. Although this method is intuitive and easy to use, such three-part decomposition might be insufficient in some cases. For example, one cannot divide the FGS for the DO term “lung small cell carcinoma” (DOID 5409) into more than three parts, while SCLC can be classified into at least four molecular subtypes (Schwendenwein, et al., 2021; Yatabe, 2020). Therefore, development of a more flexible method for dividing the correlation graphs is warranted.

Signs are derived from information in existing DBs. This inevitably introduces bias, such as the inherent incompleteness of the DBs and annotation bias; *viz*., some biological terms are associated with many genes, while others are associated with few (Gaudet and Dessimoz, 2017). To overcome this problem, one should monitor what signs are included during data processing (**Figure 1a**) and carefully tune the parameters to select reliable signs (Figure S1). Our R scripts help users perform this process.

References

Andrews, T.S., *et al.* Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc* 2021;16(1):1-9.

Aran, D., *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;20(2):163-172.

Balanis, N.G., *et al.* Pan-cancer Convergence to a Small-Cell Neuroendocrine Phenotype that Shares Susceptibilities with Hematological Malignancies. *Cancer Cell* 2019;36(1):17-34 e17.

Blondel, V.D., *et al.* Fast unfolding of communities in large networks. *J Stat Mech-Theory E* 2008:P10008.

Bodenhofer, U., Kothmeier, A. and Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* 2011;27(17):2463-2464.

Butler, A., *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;36(5):411-420.

Cancer Genome Atlas Research, N., *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* 2017;543(7645):378-384.

Cao, Y., Wang, X. and Peng, G. SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. *Front Genet* 2020;11:490.

Chen, H.J., *et al.* Generation of pulmonary neuroendocrine cells and SCLC-like tumors from human embryonic stem cells. *J Exp Med* 2019;216(3):674-687.

Chen, Z., *et al.* Ligand-receptor interaction atlas within and between tumor cells and T cells in lung adenocarcinoma. *Int J Biol Sci* 2020;16(12):2205-2219.

Coifman, R.R. and Lafon, S. Diffusion maps. *Appl Comput Harmon A* 2006;21:5-30.

Couper, P. A Student's Introduction to Geographical Thought: Theories, Philosophies, Methodologies. 2015.

Cruz, J.A. and Wishart, D.S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2007;2:59-77.

De Simone, M., Rossetti, G. and Pagani, M. Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges. *Front Immunol* 2018;9:1638.

Devitt, K., *et al.* Single-cell RNA sequencing reveals cell type-specific HPV expression in hyperplastic skin lesions. *Virology* 2019;537:14-19.

Diehl, A.D., *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* 2016;7(1):44.

Dominguez, D., *et al.* A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell Res* 2016;26(8):946-962.

Elosua-Bayes, M., *et al.* SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* 2021;49(9):e50.

Fan, J., *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 2016;13(3):241-244.

Gao, F., *et al.* DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* 2019;8(9):44.

Gaudet, P. and Dessimoz, C. Gene Ontology: Pitfalls, Biases, and Remedies. *Methods Mol Biol* 2017;1446:189-205.

Hao, Y., *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184(13):3573-3587 e3529.

Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw* 1999;10(3):626-634.

Ireland, A.S., *et al.* MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate. *Cancer Cell* 2020;38(1):60-78 e12.

Ischenko, I., *et al.* KRAS drives immune evasion in a genetic model of pancreatic cancer. *Nat Commun* 2021;12(1):1482.

Jalili, M., *et al.* Exploring the Metabolic Heterogeneity of Cancers: A Benchmark Study of Context-Specific Models. *J Pers Med* 2021;11(6).

Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28(1):27-30.

Kim, N., *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* 2020;11(1):2285.

Kiselev, V.Y., Andrews, T.S. and Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;20(5):273-282.

Kiselev, V.Y., *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;14(5):483-486.

Kristiansen, G., *et al.* CD24 is an independent prognostic marker of survival in nonsmall cell lung cancer patients. *Br J Cancer* 2003;88(2):231-236.

Kubota, S., *et al.* Dedifferentiation of neuroendocrine carcinoma of the uterine cervix in hypoxia. *Biochem Biophys Res Commun* 2020;524(2):398-404.

La Manno, G., *et al.* RNA velocity of single cells. *Nature* 2018;560(7719):494-498.

Lahnemann, D., *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* 2020;21(1):31.

Lee, J., Hyeon, D.Y. and Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* 2020;52(9):1428-1442.

Li, H., *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 2017;49(5):708-718.

Liu, X., *et al.* The reciprocal regulation between host tissue and immune cells in pancreatic ductal adenocarcinoma: new insights and therapeutic implications. *Mol Cancer* 2019;18(1):184.

Lowrance, R. and Wagner, R.A. An extension of the string-to-string correction problem. *J Assoc Comput Mach* 1975;22.

Luchini, C., *et al.* KRAS wild-type pancreatic ductal adenocarcinoma: molecular pathology and therapeutic opportunities. *J Exp Clin Cancer Res* 2020;39(1):227.

Lun, A.T., Bach, K. and Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;17:75.

Maynard, A., *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell* 2020;182(5):1232-1251 e1222.

McInnes, L. and Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. *Preprint at https://arxiv.org/abs/1802.03426* 2018.

McLeay, R.C. and Bailey, T.L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 2010;11:165.

Moncada, R., *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* 2020;38(3):333-342.

Mootha, V.K., *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;34(3):267-273.

Morrison, R.E., Baptista, R. and Marzouk, Y. Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting. *Adv Neur In* 2017;30.

Muller-Hubenthal, B., *et al.* Tumour Biology: tumour-associated inflammation versus antitumor immunity. *Anticancer Res* 2009;29(11):4795-4805.

Murtagh, F. and Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J Classif* 2014;31:274-295.

Parra, R.G., *et al.* Reconstructing complex lineage trees from scRNA-seq data using MERLoT. *Nucleic Acids Res* 2019;47(17):8961-8974.

Pasquini, G., *et al.* Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* 2021;19:961-969.

Queen, R., *et al.* Spaniel: analysis and interactive sharing of Spatial Transcriptomics data. *bioRxiv* 2019.

Reimand, J., *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* 2019;14(2):482-517.

Rempala, G.A., Seweryn, M. and Ignatowicz, L. Model for comparative analysis of antigen receptor repertoires. *J Theor Biol* 2011;269(1):1-15.

Schubert, E. and Rousseeuw, P.J. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *SISAP 2020* 2019:171-187.

Schwendenwein, A., *et al.* Molecular profiles of small cell lung cancer subtypes: therapeutic implications. *Mol Ther Oncolytics* 2021;20:470-483.

Shao, X., *et al.* scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* 2020;23(3):100882.

Stelzer, G., *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* 2016;54:1 30 31-31 30 33.

Stewart, C.A., *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* 2020;1:423-436.

Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102(43):15545-15550.

Tang, W., *et al.* bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics* 2020;36(4):1174-1181.

Trapnell, C., *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;32(4):381-386.

Villani, A.C., *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 2017;356(6335).

Wagner, F.Y., I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv* 2018.

Wooten, D.J., *et al.* Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. *PLoS Comput Biol* 2019;15(10):e1007343.

Yatabe, Y. Reassessing the SCLC Subtypes. *J Thorac Oncol* 2020;15(12):1819-1822.

Yu, G., *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010;26(7):976-978.

Yu, G., *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16(5):284-287.

Yu, G., *et al.* DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2015;31(4):608-609.

Zhang, A.W., *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 2019;16(10):1007-1015.

Zhang, W., *et al.* PIRD: Pan Immune Repertoire Database. *Bioinformatics* 2020;36(3):897-903.

Zhang, X., *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* 2019;47(D1):D721-D728.